THE PHYSICAL SOCIETY

REPORTS ON PROGRESS IN PHYSICS

VOLUME XIII (1950)

EXECUTIVE EDITOR
A. C. STICKLAND

Price to non-Fellows £2 10*s. net* (*Postage* 1*s.*)
Price to Fellows 25*s. net* (*Postage* 1*s.*)

Published by
THE PHYSICAL SOCIETY
1 Lowther Gardens, Prince Consort Road
London S.W.7
1950

PRINTED IN GREAT BRITAIN
BY
TAYLOR AND FRANCIS, LTD.,
RED LION COURT, FLEET ST.,
LONDON, E.C. 4

CONTENTS

PAGE

Preface iv

The Investigation of Eye Movements. By MARY P. LORD and W. D. WRIGHT . 1

Recent Advances in Infra-Red Solar Spectroscopy. By L. GOLDBERG . . 24

Shock Waves and the Propagation of Finite Pulses in Fluids. By W. G. PENNEY and H. H. M. PIKE 46

Ferromagnetism: Magnetization Curves. By EDMUND C. STONER . . . 83

Radio Astronomy. By M. RYLE 184

Planetary and Satellite Atmospheres. By GERARD P. KUIPER 247

Paramagnetic Relaxation Effects. By A. H. COOKE 276

Cyclic Accelerators. By J. H. FREMLIN and J. S. GOODEN 295

Mesons. By C. F. POWELL 350

PREFACE

The *Reports on Progress in Physics* have now been placed under the general supervision of the Papers Committee of the Physical Society for the selection of authors and subject matter for each volume of the Reports, and the executive editor of the publication is Dr. A. C. Stickland, Secretary-Editor of the Physical Society.

The members of the Papers Committee were as follows :

Chairman :—Professor R. E. Peierls.

Honorary Papers Secretary :—Dr. H. H. Hopkins.

Ordinary Members :—Mr. J. H. Awbery, Dr. M. Blackman, Mr. C. H. Collie, Professor G. I. Finch, Mr. H. L. Kirke, Professor L. C. Martin, Professor D. M. Newitt, Dr. W. S. Stiles, Dr. H. Wilman and Mr. C. G. Wynne.

It has been decided in future to discontinue the detailed name index given at the end of each volume since, with the Harvard system of bibliographical references now in use, references to articles are more easily found from the alphabetical list given at the end of each Report.

The cumulative name index for Volumes I–VII, which individually contain no name indexes, is published this year, and it is proposed to publish in 1951 or 1952 a cumulative index for authors and subject matter of Reports of all Volumes from I to XIV or XV.

A. C. STICKLAND,
Secretary-Editor.

THE INVESTIGATION OF EYE MOVEMENTS

By MARY P. LORD and W. D. WRIGHT
Imperial College of Science and Technology, London S.W.7

CONTENTS

	PAGE
§ 1. Introduction	1
§ 2. The eye in its orbit	4
§ 3. Classification of eye movement techniques	7
§ 4. Experimental investigations and results	9
(i) After-image experiments	9
(ii) Counting techniques	10
(iii) Direct visual observation	10
(iv) Recording by means of an attachment to the eye	13
(*a*) Mechanical recording	13
(*b*) Photographic recording	13
(*c*) Photoelectric recording	14
(v) Photography of the eyes	14
(vi) Recording of the corneal reflex	15
(*a*) Photographic recording	15
(*b*) Photoelectric recording	16
(vii) Recording of the corneal–retinal potential difference	18
§ 5. Discussion	19
Acknowledgments	22
References	22

§ 1. INTRODUCTION

THE inclusion of an article on eye movements in a volume devoted to Progress Reports in Physics requires some justification. It is in no sense a purely physical subject and has in the past claimed the attention of ophthalmologists, psychologists and physiologists rather than physicists. Moreover, since any movement of the eyes involves the neuro-muscular system, the purely physical aspects of the problem are likely to be of only secondary importance. Recently, however, physicists have shown an increasing awareness of the rôle which eye movements must play in vision and it is evident that they can make useful contributions both to the technique of recording the movements and to the interpretation of the results. Physicists have also in recent times been actively engaged on designing equipment to replace the eye in various tasks, particularly equipment for navigation at sea and in the air, and it would seem desirable that they should have some knowledge of eye movements in so far as they play a part in 'navigating' man.

The eye is frequently likened to a photographic camera, and in some respects the analogy is useful. One of the most important differences, however, is that in the case of the camera the distribution of the light-sensitive material over the photographic plate is sensibly uniform, whereas in the human eye the 'grain' of the retina is very much finer in its central area than in the periphery. Thus the fovea is populated with very narrow, closely packed cones having more or less direct connections through midget bipolar and ganglion cells to individual fibres in the optic nerve, while in the peripheral zones, the neurones are coarser and the

connections far more diffuse. Reference should be made to Polyak (1941) for the most recent exhaustive study of the histology of the retina. According to Polyak, there exists a bouquet of some 2,500 of the very narrowest cones occupying an area of about 0·1 mm. diameter in the very centre of the fovea, and it is these cones which are pre-eminently suitable for critical seeing of very fine detail. As a consequence, the examination of a particular feature of an object requires that the observer should so direct his eyes that a line passing through this foveal plateau and through the nodal point of the lens system of the eye passes through the feature under observation, Figure 1. This line defines the visual axis of the eye, and in a typical eye makes an angle of 3° or 4° with the optical axis of the lens system.

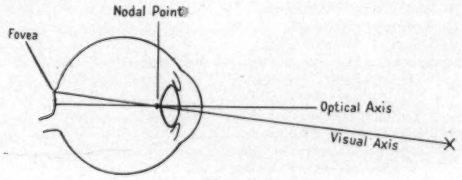


Figure 1.

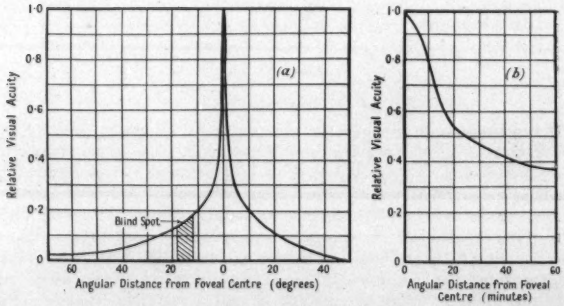


Figure 2. (*a*) Variation of visual acuity across the retina. (*b*) Variation of visual acuity within the fovea, as reported by Jones and Higgins (1947).

Subjectively, it is easy to show that the resolving power of the eye rises steeply to a maximum as the direction of regard approaches the test object. Thus Figure 2(*a*) shows the variation of visual acuity across the retina for a typical observer, while curve (*b*) shows the variation within the fovea, as reported recently by Jones and Higgins (1947). It follows that to examine an extended scene in detail an observer has to direct his visual axis successively from one point of the scene to another, hence the necessity for voluntary eye movements. Nevertheless, the usefulness of the retina is obviously not confined to the fovea, and Dodge (1907), Tinker (1946) and others have emphasized the part played by the periphery in visual perception.

Even this elementary account raises a number of interesting questions. In the first place, it is apparent that a given point on the retina cannot correspond to a given direction in space. For example, if the eyes are directed straight forward, the fovea will correspond to a direction straight ahead of the observer; if the observer looks to the right, the fovea will correspond to a direction to the right, and so on. Evidently, therefore, the spatial correlation of points on the retina is subject to constant mental adjustment, no doubt related to some extent to the tensions in the extra-ocular muscles controlling the rotation of the eye. Estimation of position in space is thus by no means a purely optical problem, and when extraneous clues of position are removed, as in the observation of an isolated point of light in an otherwise dark field, the correct location of an object may be subject to considerable ambiguity.

Again, there is the question as to how accurately the visual axis can be defined as a line in the Euclidean sense. Does the visual axis pass through one particular cone in the central fovea? If so, what distinguishes that cone from its neighbours? It is conceivable that the neural connections might become progressively less diffuse as the foveal centre was approached, until in the centre there existed one cone having a truly 'single line' connection from the retina through the optic nerve to the brain, but the histological evidence hardly supports such a suggestion. Alternatively, the cones in the centre of the central bouquet of foveal cones might possess a dominant importance by virtue of their position on the anatomical axis of symmetry of the fovea. Again, the situation might be altogether more fluid, in that at one moment one group of cones in the foveal centre might be used to fixate a point of interest in the field of view, while at another moment a different, nearby, group might be employed. In such an event, the direction of the visual axis could only be defined to within a limited accuracy of a few minutes of arc.

It is evident that the ability to direct a particular point of the fovea on to an object demands not only a fine retinal structure, but also a detailed representation of the fovea in the visual cortex. It is generally considered that the projection of the fovea occupies a relatively large area of the visual cortex (Holmes 1919, Duke-Elder 1932, Fulton 1943), although the extent to which particular parts of the brain subserve special functions is a matter of some discussion (Brodal 1948, Parsons 1949). It has at least to be remembered that an observer's ability to fixate does not depend solely on the geometry of the eye, but is an activity of the organism as a whole.

A further question concerns the steadiness of the eye. Is the retina held absolutely steady with respect to the image of a fixation object? Textbooks on ophthalmology (see, for example, Duke-Elder 1932) have suggested that small involuntary movements are continually occurring; if so, this would greatly affect the interpretation of visual experiments in which it had been assumed that a small target was under steady examination by a given group of cones, while the analysis of visual acuity data and discrimination data would become a dynamic instead of a static problem, as discussed by Marshall and Talbot (1942) and by Jones and Higgins (1947).

Another group of problems exists in connection with stereoscopic vision and stereoscopic acuity. How accurately do the two eyes move in harness together? If each eye were subject to small independent movements, how would this affect the fusion of the two images in the brain to produce a mental image in three

dimensions? Might these movements in fact play an essential part in the perception of depth, as hinted by Marshall and Talbot? The importance of these questions cannot be assessed until factual information is available about the eye movements themselves, but it is apparent that they could be fundamental to many aspects of visual perception. Indeed, electro-physiological studies of the retina (see, for example, Granit 1947) suggest that the retina is very well adapted to the detection of rapid changes in light intensity, and if this is true for normal human vision, the significance of eye movements becomes still further enhanced.

There is a further enormous field of study in the investigation of eye movements when various visual tasks are being undertaken. What happens, for example, when the eye is following a moving target? At how many points does an observer look in studying the detail of a given object? How are the movements affected by the contours of the object or by its illumination? The answers to questions of this type may have a direct bearing on the ease with which an industrial operation is carried out or with the degree of visual fatigue which it entails; in reading, for example, psychologists have made extensive studies of eye movements in relation to speed of reading, type of reading matter, the age and intelligence of the reader and so on. Reference should be made to Carmichael and Dearborn (1948) for a recent survey of this work.

Nearly all the problems referred to so far may have their pathological counterpart among observers with some degree of muscular imbalance or other abnormality. The study of eye movements is therefore of clinical interest, and may have even deeper implications related to the coordination of mind and body. The present article, however, is confined to an objective survey of the experimental techniques which have been used and to a summary of the results. Some of the material is fairly old and has already appeared in textbooks on physiological optics and ophthalmology, but in view of the current interest in the subject and of the development of new methods, a fresh assessment of the validity of the various techniques is desirable at the present time.

§2. THE EYE IN ITS ORBIT

No attempt can be made here to give an adequate summary of the anatomy of the eye and extra-ocular muscles. Reference may be made to textbooks of ophthalmology (Duke-Elder 1932, Walls 1942, Wolff 1948 etc.) for accounts of these matters, but since this is a Progress Report in Physics, it may be helpful to give sufficient information to show how the rotation of the eye differs from any simple mechanical rotation.

Figure 3 shows a view of the eye in its orbit as seen from the lateral side. The orbit is a bony, pear-shaped cavity in which the eye is held by three pairs of muscles—the four recti attached to the eye by tendons located symmetrically above and below and to left and right of the eyeball, and the two obliques attached to the lateral side of the globe and passing above and below it, Figure 4. Further support to the eye is provided by ligaments, in particular by the suspensory ligament which takes the form of a continuous band passing below the globe, while the optic nerve, the ciliary nerve and other processes are additional attachments between the eyeball and the orbit. The intervening space within the orbital cavity is occupied by the orbital fat.

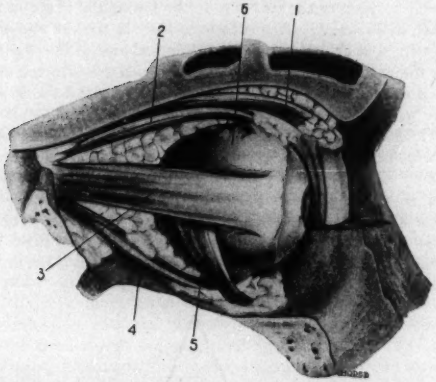


Figure 3. The extrinsic muscles of the eye seen from the lateral side.

1. Levator palpebrae superioris.
2. Superior rectus.
3. Lateral rectus.
4. Inferior rectus.
5. Inferior oblique.
6. Superior oblique.

(Reproduced from Duke-Elder's *Text-book of Ophthalmology*, Vol. I, by courtesy of Messrs. Henry Kimpton.)

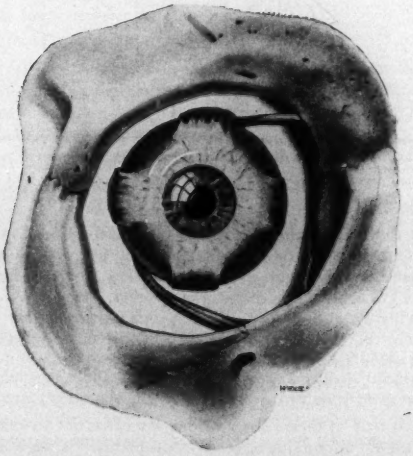


Figure 4. The oblique muscles viewed from in front.

(Reproduced from Duke-Elder's *Text-book of Ophthalmology*, Vol. I, by courtesy of Messrs. Henry Kimpton.)

The general position of the two eyes in the horizontal plane in relation to the orbits is shown in Figure 5. It will be seen that the axes of the orbits are at an angle of about 45° to each other and that the optical axes of the eyes make an angle of 22½° with the axes of their orbits when the eyes are directed straight ahead. The average diameter of the eyeball is about 24 mm., and the gap between the globe and the walls of the orbit is of the order of 3 or 4 mm.

The muscles are attached to the sclera, the outer tunic of the eyeball, by means of flat tendons whose fibres interweave with those of the sclera to merge completely with the tissue. The four recti muscles travel back from the eyeball to the apex of the orbit where they are attached to a ring, the annulus of Zinn. The superior oblique muscle is also attached to this ring, but runs close to the orbital wall before passing though a cartilaginous pulley, the trochlea, where it turns sharply backwards above the eye to its point of insertion in the eyeball. The inferior oblique is attached directly to the anterior part of the orbit.

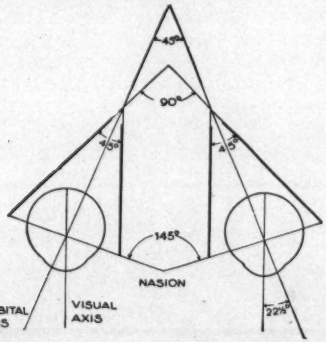


Figure 5. The location of the two eyes in the horizontal plane, in relation to the walls of the orbit.
(Reproduced from Duke-Elder's *Text-book of Ophthalmology*, Vol. I, by courtesy of Messrs. Henry Kimpton.)

According to Sherrington's principle of reciprocal innervation, contraction of one muscle is accompanied by a simultaneous and proportional relaxation of its opposing muscle. This suggests a relatively simple arrangement by which a rotation of the eye could be produced, but in practice any movement of the eyes involves a highly complex co-ordination of all twelve muscles. Moreover, the proximity of the centre of rotation of the eyeball to its geometrical centre will in part be determined by this co-ordination, in addition to the passive restraint exercised by the relatively flexible fat and other tissues in contact with the eyeball.

The length of the recti muscles is about 40 mm. and they can contract by about one third to one half of their length. This would permit an angular rotation of the eye of about 60° to one side or other of the primary (straight ahead) position, although Lancaster (1941) estimates that about 99% of eye movements involve rotations of the eyes themselves of less than 15°, the head being turned or tilted when objects at greater angles than this from the centre of the field of view are to be examined. Lancaster also concludes that not more than 5%

of all the fibres in the muscles need to be contracted at any one time to produce a force sufficient to overcome frictional forces and to rotate the eye at the observed velocity. This result assumes the mass of the eyeball to be about 8 gm. and the maximum tension which could be exerted in each of the muscles to be about 750–1,000 gm. wt. if all the fibres contracted at once.

It is impossible in the space of this article to give any useful summary of the prevailing ideas concerning the neuro-muscular processes by which contraction of the muscles is set into operation and controlled. It should, however, be noted that the extra-ocular muscle fibres are the finest of all the voluntary muscles in the body (Duke-Elder 1932), that there is an unusually high proportion of motor nerve fibres to muscle fibres (Fulton 1943), and that the sensory nerve fibres are also very numerous (Duke-Elder 1932). These features might all be expected to contribute to the very accurate and rapid movements required of the eyes and to the postural and positional importance of these movements. It has also to be remembered that eye movements cannot be considered in isolation, but are co-ordinated with other muscular movements, more especially those of the head. Thus, when the head moves, reflex compensatory movements of the eyes take place, and, to quote Lancaster, "the innervation of the muscles moving the head is an integral part of the neuro-muscular mechanism for ocular movements".

§ 3. CLASSIFICATION OF EYE-MOVEMENT TECHNIQUES

Eye movements have been studied by both subjective and objective methods. The subjective methods may be divided into

(i) after-image experiments,
(ii) counting techniques,

in which, in (i), the subject observes the movement of an after-image over the visual field, while in (ii) he determines the angular separation of the finest pattern which he can just count.

Objective methods can conveniently be classified as follows:

(iii) direct visual observation of the eyeball,
(iv) recording by means of an attachment to the eye,
(v) photography of the eyes,
(vi) recording of the corneal reflex,
(vii) recording of the corneal–retinal potential difference.

In so far as the results of these objective methods are used to make deductions about the movement of the image across the retina, it is necessary to make certain assumptions about the behaviour of the eyes as described briefly in § 5.

In (iii), the eyes of the subject may be observed either directly or through an optical instrument such as a telescope or microscope. In (iv), mechanical, photographic and photoelectric recording systems have all been employed. Thus, in some investigations, a recording lever has been placed in mechanical contact with the eye. In photographic recording, some workers have taken intermittent photographs of a speck of white pigment placed on the cornea, while others have photographed the reflection from a mirror attached to the eye, on a continuously moving film. Photoelectric recording has been used in conjunction with a lever in contact with the eye, a screen attached to the lever being arranged to modulate

the intensity of a beam of light striking a photocell. The eye movements could thus be deduced from the amplified photocurrents as recorded with a galvanometer oscillograph.

(v) simply requires photography of the eyes on intermittent or continuously moving film, the eye movements being deduced from the displacement, in successive pictures, of structural features of the eyes.

In (vi), use is made of the fact (*a*) that the corneal surface reflects a fraction of any light incident on it in much the same way as the convex surface of a lens, with a radius of curvature of the surface of about 8 mm., and (*b*) that the centre of curvature of the cornea is separated from the centre of rotation of the eye by about 5·3 mm. As a consequence, the angle at which a beam of light is reflected from the cornea changes as the eye rotates, while the virtual image (the corneal reflex) of a light source formed by the cornea is displaced (Figure 6). These

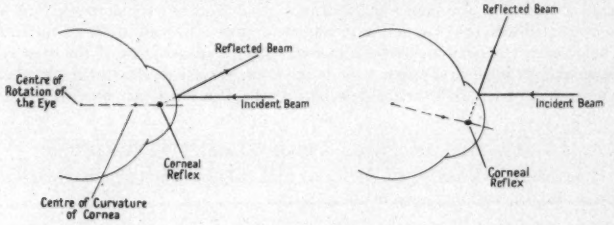


Figure 6.

movements have been recorded both with photographic and with photoelectric systems. Thus the corneal reflex may be photographed on either stationary or moving film and the eye movements deduced from the displacements of the image, while in photoelectric recording the principle is similar to that described under (iv), except that the light modulation is produced as the beam of light reflected from the cornea moves over a straight edge or neutral density wedge.

The origin of the corneal–retinal potential difference used in (vii) is still a matter of some doubt, but a small potential difference is found to exist between the cornea and retina which probably originates from the differing metabolic rates of the various regions of the eye. When the eye turns to right or left, a change occurs in the potential difference between points of the skin on either side of the eye, produced by the corneal–retinal potential, and a similar change occurs between points above and below the eye when the eye rotates up or down. These potential changes can be picked up by pairs of electrodes attached to appropriate points of the skin, and when amplified and recorded, can be used to deduce the horizontal and vertical deflections of the eye. It is claimed that this method yields eye movement records which are independent of any head movements, apart, of course, from the reflex compensatory eye movements associated with head movements.

An additional method which does not appear to have been used would be to illuminate the retina and then record the movement of the image of the fundus formed by the optical system of the subject's eye. In that case, assuming that the magnification produced by the eye can be considered constant, the recorded

movement would be directly proportional to the movement across the retina of the retinal image of the fixation object. With the recording systems at present available, this method would, however, require a high level of illumination on the subject's retina in order to obtain reasonable sensitivity. This is probably the reason why the method has not been employed.

§4. EXPERIMENTAL INVESTIGATIONS AND RESULTS

(i) *After-Image Experiments*

After-images have been widely used by many workers to study eye movements (Helmholtz 1925, Duke-Elder 1932), and it is only necessary to refer here to typical experiments. Fixation movements, for example, have been studied by Dodge (1907). He projected the after-image of an illuminated wedge on to a piece of graph paper and found that in spite of the utmost efforts to fixate a point on the paper, the after-image would not remain stationary. Dodge estimated that the amount of wandering was about 10 minutes of arc, spasmodically interrupted by excursions of about $\frac{1}{2}°$. Both types of movement were found to increase with fatigue.

Dodge has criticized the after-image technique to the extent that the after-image may itself influence the ability to fixate, and where an unsymmetrical after-image such as a wedge is used, there may be a tendency to drift. This latter objection can be overcome by using an illuminated annulus to form a ring-shaped after-image with the fixation target in the centre, but it is found that the wanderings of the after-image ring relative to the fixation target are still of the same order of magnitude as those reported by Dodge.

Other difficulties of the after-image method include the limited duration of the after-image, which prevents the movements being observed for an extended period of time, and the lack of sharpness of the image, which limits the accuracy of observation to a few minutes of arc. The same limitations apply to the subjective observations of the fovea and macula as seen by viewing a white area of light through a colour filter transmitting the extreme violet and red ends of the spectrum. The foveal centre usually appears as a small reddish pip surrounded by a further reddish zone, and the accuracy with which the pip can be directed on to a small fixation spot is of the order of 10 minutes of arc or less. Here again, however, the image fades rapidly, but can be renewed by removing the colour filter momentarily.

Rapid excursions of the eye may be missed in the after-image observation because of the relative faintness of the image. This objection is overcome in the type of observation in which clearly defined areas of retinal adaptation are produced by gazing for a minute or two at a fixed point in a pattern of bright and dark areas. When the corresponding areas of local adaptation have been built up, any sudden excursion of the eye will be revealed by the appearance of very bright or very dark edges round the pattern.

Landolt (1891) in an early investigation employed after-images to investigate the nature of a voluntary sweep of the eyes through a 30° arc. A magnesium lamp was placed at the centre of the arc and after the eye had swept from one end of the arc to the other, it was found that the after-image trail was interrupted by knots of light. Landolt concluded that the sweep of the eye was not continuous but saccadic and that the knots corresponded to the fixation pauses.

The speed of eye movements was studied as long ago as 1869 by Lamansky. An intermittent source of light was located between two fixation points and, as the eye swept from one fixation point to the other, a succession of after-images was impressed on the retina; these could be counted by an experienced observer and the time of traverse calculated. This type of experiment has been developed and elaborated by a number of other workers (see, for example, Cobb and Moss 1925).

O'Brien and Dickerman (1948) have recently reported an experiment in which a point of light was viewed in a dark circle with a light surround. The star was illuminated by flashes having a frequency of 18–240 per second and was exposed to the observer for short bursts ranging from 0·2 to 0·01 second. If involuntary fixation movements occur, O'Brien considered that the point of light should appear as a series of multiple stars at various positions within the dark circle, but no such effect was observed. The conclusion that steady fixation was being maintained presupposes that any eye movement would have a different effect on the apparent position of the intermittent light source compared with that of the permanent surround. This assumption requires independent confirmation.

Torsion of the eyeball has also been investigated by after-image techniques in which, for example, an illuminated vertical cross has been viewed for a few seconds and its after-image subsequently observed for various positions of the eyes, preferably as seen projected against a curved surface to which the visual axis is always perpendicular. Horizontal and vertical movements of the eyes produce no rotation of the after-image, whereas if the eyes are moved obliquely, the cross takes up a sloping position, indicating some torsion of the eyes. Loring (1915), however, using an improved technique, considered that some torsion also occurs even for vertical and horizontal eye movements.

These after-image methods are discussed extensively in ophthalmological textbooks and have provided the classical observations upon which knowledge of eye movements has been founded.

(ii) *Counting Techniques*

An experiment by Landolt (1891) exemplifies the counting technique. He determined what he considered was the smallest angle through which the eye could move voluntarily by viewing a recurring pattern at various distances. The separation between the patterns was 13 mm., and he increased the distance of view until the number of patterns could only just be counted. The angular separation between the patterns was then found to be about 5 minutes of arc.

The physiological significance of this angle is a matter of some doubt, but it seems reasonable to suppose that it must be related fairly directly to the accuracy with which the direction of the visual axis can be defined. It might therefore be a measure of the fixation area of the fovea or perhaps an indication of the magnitude of the random eye movements occurring during fixation.

(iii) *Direct Visual Observation*

Javal (1879) made a systematic study of the saccadic movements of the eyes by direct observation, but to minimize the distraction of the experiment he placed a mirror in front of the subject so that the eye movements could be observed from behind the subject by reflection in the mirror. Such experiments have the merit of simplicity and have been extensively employed, but small movements cannot be observed without some optical magnification. Thus a rotation of the eye of

1° produces a movement of only 0·2 mm. round the circumference of the eyeball. In one instrument (Newhall 1928), a 20-dioptre convex lens was used as a simple magnifier to increase the sensitivity.

Of more importance from the point of view of this Report are the fairly elaborate optical systems which have been used to locate the centre of rotation of the eye for horizontal movements, to investigate the behaviour of the corneal reflex and to observe the behaviour of the eyes when using a stereoscope. In the ocular kinemometer of George, Toren and Lowell (1923), a sighting tube was mounted on a horizontal arc graduated over 180° with the axis of the tube pointing to the centre of the arc, while the arc could be accurately moved in three mutually perpendicular directions. A microscope was mounted with its axis at right angles to the axis of the sighting tube and intersecting the latter at a point about 15 mm. from the centre of the arc. The subject's head was fixed by means of a forehead rest and band and wooden mouthpiece, and in such a position that the centre of his eye was approximately coincident with the centre of rotation of the sighting tube. With the subject looking straight ahead, the sighting tube was finely adjusted to bring its axis in line with the visual axis of the subject, as judged by the alignment of a slit and needle fixed at either end of the sighting tube. The observation was repeated for different positions of the sighting tube round the arc, and a position was found for the centre of the arc such that no readjustment was required to secure alignment of the visual axis and the sighting tube as the latter was moved round the arc. That such a position could be found implied that the visual axis for the different fixation directions in the horizontal plane passed through a common point which has been termed the visual centre, and that this point was coincident with the centre of the arc. However, observation of the corneal apex through the microscope revealed a movement of the apex along the visual axis for different directions of the axis, showing that the eye did not rotate about a point on the visual axis. Using a geometrical method, George, Toren and Lowell showed that for the range 20° templeward to 30° nasalward, the centre of rotation of the eye could be considered to be a fixed point situated 15·4 mm. from the corneal vertex and 1·65 mm. nasalward from the visual axis. For angles outside this range the centre was no longer fixed.

Park and Park (1933) repeated these observations using a different sighting target and a photographic method of recording the position of the cornea. These workers concluded that the centre of rotation of the eye was not a fixed point either in the eye or relative to the head, and developed an interesting analysis of the motion of the eye in terms of space and body centrodes and the instantaneous centre of rotation. The positions of the centre of rotation for three fixation directions as found by Park and Park were as follows:

Direction of visual axis	39° nasalward	4° nasalward	38° templeward
Distance of centre of rotation from corneal vertex (mm.)	14·73	13·92	12·95
Perpendicular distance of centre of rotation from axis (mm.)	1·066	1·653	0·893

These results are of very great interest, but it should be realized that the data are obtained from measurements on an eye viewing a stationary target in each

fixation direction. It cannot be taken as axiomatic that the centre of rotation when the eye is actually in motion coincides with the instantaneous centre of rotation as derived by Park and Park, especially in the case of very small rotations. This point is discussed again in § 5.

For various conditions of observation, Park (1936 b) has investigated the behaviour of the line through the visual centre and the centre of curvature of the cornea, relative to the visual axis. For this he developed the anglometer (Park 1936 a) which is in many respects similar to the kinemometer; the main difference is that the sighting tube is replaced by a telescope, and the microscope is dispensed with. Coincidence of the visual axis of the subject with the axis of the telescope is obtained by lining up a fixation target whose virtual image is located in the focal plane of the telescope with an annulus placed in front of the telescope objective. By means of a supplementary target, the subject then maintains the direction of his visual axis, while the observer rotates and focuses the telescope centrally on to the virtual image of the fixation target as seen by reflection in the corneal surface. The telescope axis must then pass through the centre of curvature of the cornea and, by virtue of the dimensions of the apparatus, through the visual centre for the average eye. The angle between the two settings of the telescope could be measured to an accuracy of about 15 minutes of arc.

Results reported by Park (1936 b) and by Park and Park (1940) showed considerable variation in the behaviour of the corneal reflex from subject to subject: some maintained the reflex steady for several minutes and then a new position would be taken up; with others, after an initial steady period, there would be a drift in one direction or another, followed by a reversal; for other subjects there were almost continuous random movements, interrupted by short pauses. The effect of eserine, homatropine and atropine were also studied, and in general these drugs tended to reduce the angle between the visual axis and the corneal reflex, frequently to zero, although there were many exceptions which showed an increase in angle. With aphakic observers, the visual axis and the corneal reflex were found to coincide without exception, although, on account of the low visual acuity, the accuracy of setting was reduced. Reference is made in § 5 to theoretical suggestions put forward by Park and Park to explain these results.

Peckham (1934) constructed a large stereoscope with which the subject was required to fuse two identical pictures placed in each half of the instrument. Prisms, either base-in or base-out, were then placed in front of the subject's eyes and he was again asked to fuse the pictures. The movements of his eyes were observed through two telescopes mounted behind the stereoscope and Peckham reported that when overcoming, for example, a 10° prism, the eyes of the subject would often move through an angle of less than half this magnitude. Peckham therefore concluded that although fusion was maintained, the two foveas were fixating on points in the visual field 3° or 4° apart.

Reference should also be made to a rather different binocular observation in which two diagrams are presented in a stereoscope to the right and left eyes, the diagrams being similar except that there are one or two additional details in each view which are not present in the other. Fusion of the similar parts of the views is easily achieved, but it is found that some change in convergence can occur, as indicated by a change in the apparent separation of the dissimilar details, without affecting the fusion of the major part of the view. This effect has been observed both in foveal and extrafoveal vision and has been described as 'fixation disparity' by Ogle, Mussey and Prangen (1949). Its greatest magnitude is about $\frac{1}{2}°$.

(iv) *Recording by means of an Attachment to the Eye*

(*a*) *Mechanical Recording.* In mechanical recording, use is generally made of the protuberance of the cornea. Thus in the method described by Ohm (1914, 1916, 1928) a lever, attached at one end of the eyelid, rested against the cocainized conjunctiva near the outer cornea and, as the eye moved, the lever was raised or lowered and the movement could be recorded on a revolving drum. As Cords (1927) has pointed out, useful results can only be obtained if the lever is firmly attached to the eye and if pressure or pulling on the eye is avoided.

Various workers, as reviewed by Carmichael and Dearborn (1948), have employed capsules laid over the eyeball; movement of the corneal bulge under the capsule then produces a change in air pressure which has been recorded pneumatically. Other workers have used light-weight cups fitting over the cornea itself with an aperture or window in the centre to allow the subject to see. Recording has then been made by various fittings attached to the cup: thus Delabarre (1898) embedded a wire ring in a plaster cup and attached one end of a thread to the ring and the other end to a recording lever, Huey (1898, 1900) attached a lever directly to a plaster cup, while Orchansky (1899) used an aluminium cup with either a mirror to reflect a light beam on to a projection screen, or a recording lever.

These mechanical methods were adequate to study some of the grosser eye movements and to show, for example, the saccadic nature of reading movements. They must, however, have proved extremely inconvenient to use and had only a limited sensitivity.

(*b*) *Photographic Recording.* Photographic recording provides greater scope, and Judd, McAllister and Steele (1905) achieved considerable success by taking intermittent photographs (about 9 photographs per second) of the eyes and part of the face, a flake of Chinese white having been attached to each cornea. The subject bit on a piece of wood and his head was supported at the back and sides, and in order to allow for any residual head movements, the reflection from two small steel balls attached to a spectacle frame and worn by the subject, were also photographed. The eye movements were then deduced from the shift of the Chinese white image on the film relative to the steel ball image, and it was claimed that by this means the maximum error of measurement did not exceed $7\frac{1}{2}$ minutes of arc. Experiments on steady fixation showed that movements of the eyes of at least 30 minutes took place. The use of intermittent photography at such slow speeds necessarily involved considerable gaps in time in the records and a relatively coarse recording of the movement. A much greater refinement of the technique would now be possible using modern photographic equipment.

Several workers have photographed the reflection from a mirror attached to the eye. Thus Adler and Fliegelman (1934) anaesthetized the right eye of the subject and placed a small plane mirror (3 mm. diameter) on the temporal limbus. The subject was seated with the head in a chin and forehead rest and, to prevent blinking, the upper lid was supported by the middle finger of the subject's left hand. Light from a vertical slit was reflected in the mirror and focused on a film which moved continuously in the vertical direction. This gave a continuous record of the horizontal movements of the mirror, a displacement of about 0·4 mm. in the track on the film corresponding to a rotation of the eye of 1 minute of arc. Monocular and binocular fixation of various types of test object under both steady and intermittent illumination were investigated. In all cases small movements of the order of 2 minutes of arc at a frequency of about 100 per second

were recorded; they were superimposed on larger movements whose nature depended on the conditions of observation.

Marx and Trendelenburg (1911) used an aluminium corneal cup and mirror somewhat similar to Orchansky's, but the light beam was reflected by the mirror into a camera and focused on a continuously moving light-sensitive paper. Dohlman (1925) used a similar method except that the aluminium cup was replaced by a rubber cap.

In these earlier experiments it was generally necessary to anaesthetize the cornea, but with current experience of contact lenses, it should now be possible to employ the same principle under more favourable conditions. A modern development on these lines has been described recently by Riggs and Ratliff (1949) in which a contact lens was fitted over the subject's eye. A plane mirror was attached to the contact lens, and recording of both vertical and horizontal movements during monocular fixation was obtained photographically. A very small eye 'flutter' was observed having a magnitude of about 18 seconds of arc and a frequency of 30–70 cycles per second. There were also slower excursions of several minutes of arc and of irregular frequency, together with slow drifts and compensatory jumps.

Several workers have photographed eye movements with the lids closed. For example, Dodge (1921) lightly pressed a mirror against the lid over the cornea and recorded the deflection of a light beam reflected from the mirror; as the eye moved, the mirror tended to take up a position tangential to the corneal surface. The method has been applied to the study of nystagmus.

Rodin and Newell (1934) have devised a method of photographing eye movements when the eyes are bandaged. Small pieces of spring were pinched on to the sclera of the anaesthetized eyes and an x-ray picture taken through the bandages.

(*c*) *Photoelectric Recording.* Dohlman (1935) appears to have been the only worker to combine the use of an attachment to the eye with a photoelectric method of recording the eye movements. The eye under investigation was cocainized and the lids held apart with a blepharostat; a rubber cap was then applied to the cornea, to which it adhered by suction. An aluminium screen was connected to the cap by a system of levers with rubber connections and so positioned that it half intercepted a beam of light incident on a photocell, both light source and photocell being attached to the head by a head-band. Movement of the eye then modulated the light entering the cell and the amplified photocurrent was recorded with a galvanometer oscillograph. Side-tc-side movements only were recorded. Dohlman applied the method to the determination of the smallest amount of acceleration or deceleration that would produce nystagmus in a subject seated in a rotating chair in order to obtain a criterion of the performance of the vestibular apparatus. Very few results are, however, given.

(v) *Photography of the Eyes*

Dodge and Cline (1901) appear to have been the first to photograph the eyes in motion. An image of the sclera, the iris and the pupils was focused on a 5 in. by 7 in. falling plate and a record obtained of alternating fixations to right and left, and also of reading eye movements.

With the development of modern kinematograph equipment, the convenience of this method has been greatly enhanced, although, since the sharpness with which any particular feature of the eye can be defined on the film is limited, it

can only be applied to the study of the grosser movements of the eyes. Karslake (1940) has described a modern development of the method which he has applied to a study of the attention value of advertisements. The film camera was located above the subject's head and a large half-silvered mirror mounted between the subject and the reading matter at such an angle that an image of the subject's face was reflected into the camera. It is, of course, unnecessary to fix the observer's head in this method. Good agreement was found between the subject's report of his fixation pattern and the movements as deduced from a frame-by-frame examination of the film.

(vi) *Recording of the Corneal Reflex*

(*a*) *Photographic Recording.* Photography of the corneal reflex is probably the most popular method of recording eye movements. It was first used at the beginning of the century by Dodge and Cline (1901) and by Stratton (1902, 1906). In the form of the apparatus described by Dodge in 1907, light from an A.C. operated arc lamp was passed through a blue glass filter and a vertical slit, and was reflected axially into one eye by a glass plate. The image formed by reflection in the cornea was then photographed through a horizontal slit on to a falling plate. A record of the side-to-side movement of the corneal reflex was thus obtained and, since the arc was A.C. operated, the record was broken and gave the time scale also. The subject was seated with his head supported by an upper jaw rest, a wooden nose bridge and a mastoid rest. The lateral position of the head was also recorded on the plate as in the work of Judd, McAllister and Steele (1905).

Dodge applied the method to the measurement of fixation eye movements, but quotes very few results. He also photographed the reflex during following movements using a moving pendulum as the target, but the amplitude of the pendulum swing is not stated. He found typical corrective movements at the commencement of all following movements, but subsequently the eye settled down to a generally smooth pursuit movement. One or two records of reading eye movements are also given.

Stratton (1902, 1906) photographed the corneal reflex on a stationary plate to study the behaviour of the eyes when viewing a diagram; when the subject followed round the circumference of a circle, for example, the eye followed a path that was more triangular than circular and that included a number of fixation pauses *en route*. He also recorded eye movements when the gaze was voluntarily transferred from one fixation point to another for different directions of separation. He found that the eye rarely followed the shortest path between the two points and that there was sometimes overshooting; the overshooting indicated that the extra-ocular muscles on the nasal side had too strong an action. The horizontal movements were the most direct and the most consistent.

There have been many subsequent developments of corneal reflex photography as described in detail by Carmichael and Dearborn (1948). These include the large 70 in. camera described by Tinker (1931) and the Harvard Optokinetograph developed by Comfort and Dearborn (Taylor 1937), each instrument recording the horizontal movements of both eyes and being especially suitable for studying reading movements.

The need for a more portable instrument has been met by the Ophthalmograph designed by Taylor (1937) and made by the American Optical Company. A small lamp (automobile headlight) is used as a light source for each eye and the corneal

image of the source is focused on to 35 mm. film moving vertically at constant speed. The subject's head is held with head and temple clamps and a chin rest, and the reading matter placed on a stand at the normal reading distance. The horizontal eye movements are then shown by the lateral displacements of the light track on the film, and although the sensitivity of the instrument is not more than 1° or 2°, the saccadic nature of the movement in reading, the fixation pauses and the regressions are all clearly indicated.

Apparatus to record both vertical and horizontal movements by photographing the corneal reflex is necessarily more complicated but, as in the instrument by Weaver (1931), it can be accomplished by using both vertically and horizontally moving films. However, by introducing a totally reflecting prism into the optical system, Clark (1934) arranged that the motion of the beam recording the vertical movement was rotated through 90°, so that both vertical and horizontal movements could be recorded by tracks showing horizontal displacements on a single film moving vertically. Records obtained by Clark (1936) while the subject used a stereoscope indicated random and irregular eye movements during fixation, and Clark concluded that any retinal position within an area of 1–1·5° from the centre of the fovea may be used to fixate a point.

A recent development of the corneal reflex method has been described by Hartridge and Thomson (1948) in which, in an attempt to avoid artifacts due to head movements, the light source, fixation target and essential optical components were attached to a plaster of Paris hat which fitted on the subject's head. The corneal reflex was imaged in the same plane as a small reference source, and both reference source and reflex were photographed together by a ciné camera mounted independently on a bench and taking 60 frames per second. From the sharpness of the corneal image on the film, it was concluded that the eye could not be performing rapid vibratory movements, although no figures for the sensitivity of such a judgement are given. It was found that the plaster hat did not completely eliminate movements of the apparatus relative to the head, and plans for an improved apparatus were described in which the subject bites on a dental plate attached to an optical system which is counterpoised and free to rotate in all directions. No constraint would therefore be imposed on head movements, and the fixation target would be fixed to the wall beyond the apparatus and not form part of the main optical system, as in the first apparatus.

The study of eye movements in reading is one of the most important applications of the photography of the corneal reflex. Tinker (1946) has given a comprehensive survey of the results so far obtained. He states that the duration of the fixation pause varies from 0·217 to 0·404 second depending on the type of material read, and that some 5–10% of the time during reading is taken up by eye movements. He also states that the duration of the fixation pauses increases as the intensity of illumination decreases and that the eye movement pattern is influenced by the visibility, word form and line length of the material. Tinker considers that the central processes of perception and assimilation are of major importance in reading and that the eye movements are an expression of these processes. He therefore deprecates the school which advocates the improvement of reading by eye movement training. In common with Dodge (1907), he emphasizes the part played by peripheral vision in reading.

(*b*) *Photoelectric Recording*. A photoelectric method of recording the corneal reflex movements has been described by Lord (1948), in which rotations of the

eye as small as 1 minute of arc can be detected. In this method the head movements have been greatly reduced by the subject lying on his back on the floor with his head resting on a firm foundation of variable height, and partially supported by a sling; he also bites on a dental impression fixed to an adjustable mounting. An ultra-violet beam of wavelength 3650 A. is reflected into the subject's eye in such a direction that it falls on the blind spot when the subject is viewing the fixation target, so that he is hardly aware of the ultra-violet patch. The part of the ultra-violet beam reflected at the corneal surface is incident on a half-aluminized mirror, where it is divided so that one part falls on a horizontal straight edge and the other on a vertical one. Thus as the eye moves, more or less light passes each straight-edge and the variable light flux is focused on to two electron multiplier photocells. The output of each multiplier is passed to a single stage balanced D.C. amplifier and thence to the vertical deflecting plates of a cathode-ray oscillograph. The traces on the two tubes are photographed one above the other on a continuously moving film travelling in the horizontal direction. Dots at 1/50 second intervals along the middle of the film are obtained by focusing the image of a pinhole, illuminated by a discharge lamp in series with a half-wave rectifier, on to the film.

The side-to-side and up-and-down movements of the eye are given by the two tracks in which, under conditions of maximum sensitivity, a displacement of about 1 mm. corresponds to an eye rotation of 1 minute. Calibration of the system was carried out with the aid of a lens surface of the same radius as the cornea; this could be tilted about a point 5·3 mm. below its centre of curvature, corresponding to the assumed distance between the centre of curvature of the cornea and the centre of rotation of the eye.

In spite of the precautions taken to keep the head fixed, a very small amount of residual head movement remained. This, however, was recorded with the artificial glass cornea mounted on an aluminium strip and attached to the teeth of the subject's upper jaw by a dental regulation band; the movements proved to be periodic at a frequency corresponding to the heart beat (Lord and Wright 1948). Allowance could then be made for these head movements when analysing the records from the live cornea.

Under the steadiest conditions of fixation, Lord and Wright found that for periods of about 0·5 second the eye appeared to be steady, but that these steady periods would be interrupted by a rapid flick corresponding to an angular movement of the eye ranging from 3–14 minutes of arc and lasting 0·02–0·03 second. The fixation direction after each flick might then change by a few minutes. Little evidence of the high-frequency movements reported by Adler and Fliegelman (1934) was found, although on many occasions fixation was very unsteady and gave records showing movements of 20 or 30 minutes of arc or more.

In a modified apparatus (Lord and Wright 1949) the straight edges were replaced by 1 in. square neutral density wedges to enable larger excursions of the eye to be recorded, although at reduced sensitivity. Records have thus been obtained when the direction of regard has been consciously shifted backwards and forwards between two fixation points for a separation of the points in the range 1–4°, and when following a fixation point performing either vertical or horizontal oscillations of similar magnitude. With rapid flicking, the change in fixation was generally accomplished in a single jump, although there was sometimes some overshooting. For the same angular separation of the fixation points the

recorded movement generally tended to be smaller in the vertical than in the horizontal direction, suggesting possibly a different position for the centre of rotation. A slow, and, to the subject, apparently smooth, transfer of fixation from one point to another was shown by the record to be saccadic in nature, but a smooth pursuit movement was found when following a moving target, in agreement with Dodge (1907). An unpublished record obtained when looking round the circumference of a circle showed the movement to be saccadic, in agreement with Stratton's findings (1906). The results have generally been reasonably consistent except that in the early records for vertical flicking for subject M.P.L. only a momentary change in fixation was recorded, whereas in later records the change in direction was maintained until the next flick took place.

(vii) *Recording of the Corneal–Retinal Potential Difference*

Although the corneal reflex method has been widely used for studying reading movements, the rather artificial conditions under which the reading has to be carried out is a considerable disadvantage, and continuous records over a long period of time are difficult. The possibility of recording by means of the potential changes which occur around the eye when the eye moves has therefore been explored by a number of workers and has been shown, notably by the large-scale experiments of Carmichael and Dearborn (1948), to be an eminently practical method where relatively coarse movements only have to be studied.

Schott (1922) made early experiments using as electrodes spirals of copper wire which he attached to a spectacle frame so that they pressed against the cocainized eye. Meyers (1929) eliminated the need for an anaesthetic by attaching electrodes to the temples at both sides of the head. It was found that maximum deflection was obtained when the line joining the electrodes was parallel to the plane of the eye movement; no deflection was obtained when the line was perpendicular to this plane. Jacobson (1930 a, b) attached electrodes of platinum foil, wrapped in cotton and soaked in saline solution, to the orbital ridge and to the mastoid bone behind the ear, and obtained signals which clearly differentiated the direction in which the observer was looking. He regarded the signals as action potentials from the extra-ocular muscles.

Mowrer, Ruch and Miller (1936), however, have pointed out that the polarity of the signals obtained when electrodes are placed on each temple is the reverse from that expected on the action potential theory, but is in agreement with a corneal–retinal origin of the potential difference. These workers obtained direct confirmation of the existence of a difference in potential between the front and back of the eye by exploring the freshly excised eye of the turtle; they found that the cornea was always positive with respect to the back of the eye.

Miles (1939 a, b, c, 1940) investigated the effect of the experimental conditions on the potential difference. Closing the eyelids and pressing on the eyes produced a slight increase in the potential, while if the subject were in a relaxed position the potential gradually decreased. Light adaptation produced an increase, and dark adaptation a decrease, in the potential difference.

The potential changes that have to be recorded are generally less than a millivolt, so that successful recording presents difficulties if the result is not to be upset by changes in the contact resistance between the electrode and the skin. Special attention has therefore to be paid to the electrodes, and either a valve voltmeter or an amplifier and oscillograph may be used for recording. In Carmichael and Dearborn's experiments (1948), an amplifier and ink-writing

oscillograph have been used, and have been found advantageous both in obtaining multiple records of long duration and from the point of view of cost. Their electrodes were made of chlorided silver and were shaped like miniature hats with crowns and brims. Before attaching the electrode the skin was cleaned with ether to remove oil and dirt; the brim of the electrode was then attached to the skin with collodion, while an electrode jelly was inserted under the crown of the electrode to provide good contact.

The chief items to be recorded in a study of reading movements are the number of fixation pauses per line, the number of regressions per line and the number of lines read per unit of time. All this information can be obtained from a record of the horizontal movements of the eyes alone so that, in Carmichael and Dearborn's experiments, the main eye movement records were obtained from two electrodes only, which were attached to the face at the external canthus of each eye. However, the frequency of blinking is also of some interest in these problems, so that two further electrodes were attached above and below the right eye to record the upward movement of the eye associated with blinking.

Their experiments were concerned with the correlation of eye movement with the development of fatigue and with the comprehension of the material being read during prolonged reading, and the investigation embraced different types of reading matter and subjects of varying mental ability and education. A detailed review of the experiments themselves and of the results is outside the scope of this Report, but there is no doubt that the technique of the corneal–retinal potential is very suitable for this type of work because of the relative freedom with which the subject can move his head and make himself comfortable, and the more natural conditions under which the reading material can be located and illuminated. It should be noted that the validity of the method depends on a comparison with more direct observations of eye movements, amounting, in effect, to a calibration in terms of an optical method.

For the study of small eye movements, however, the possibilities of the method appear to be limited. Baudouin, Fischgold, Caussé and Lerique (1939) consider that the action potentials of the extra-ocular muscles would be sufficiently large to complicate the interpretation of the records of small movements. There does not, in fact, appear to be any account of the application of the corneal–retinal potential difference method to the investigation of fixation eye movements.

§ 5. DISCUSSION

In selecting a technique of measuring eye movements, an important consideration is that the technique should have adequate sensitivity for the purpose in view. Thus, in the investigation of fixation eye movements, the highest sensitivity is required, and from this point of view the mirror attachment method devised by Adler and Fliegelman (1934), the photoelectric corneal reflex method devised by Lord (1948) and the contact lens method devised by Riggs and Ratliff (1949), in which movements of a minute of arc or less can be detected, have proved superior to other methods. For the purpose of investigating reading eye movements, a sensitivity of the order of a degree may be adequate, and the photographic corneal reflex and the corneal–retinal potential methods have proved suitable. To obtain general information about the movements of the eyes, subjective methods, methods of direct photography and of direct visual observation have sufficient sensitivity.

Apart from the question of sensitivity, the value of a technique is greater the smaller the amount of interference it produces in the natural condition of the eye. From this point of view the after-image techniques are excellent provided the intensity of the source giving rise to the after-image is not too great. Direct visual observation is also good so long as the subject is not unduly disturbed by the proximity of the operator. Lack of interference is also one of the advantages of the corneal–retinal potential method, and of direct photography of the eyes, providing the amount of light entering the eye is kept low. Successful photography of the eye structures, however, probably requires a higher illumination than photography of the corneal reflex. The worst methods with respect to interference are those in which there is an attachment to the eye, especially if the eye has to be anaesthetized. It may also happen that the movements of the eye may be affected by the attachment; thus Dodge and Cline (1901) reported that the times required for the eye to move through various angles when measured by the corneal reflex method were significantly shorter than those determined by Huey (1900) with his plaster-cast attachment to the cornea. Where systematic testing of a large group of subjects is envisaged, the general inconvenience of having to fit an attachment to the eye for each subject would probably make such a method impracticable.

In addition to minimizing interference with the eyes, it is, of course, desirable not to place the subject himself under unnatural constraints. From this point of view, all methods in which the head is fixed are at a disadvantage. It is unnecessary to fix the head in the subjective methods, the corneal–retinal potential method and sometimes in the direct photography of the eyes. For most other methods there appears to be no practical alternative to the use of some head-fixing device. It is probably less of a strain to have the head fixed with the observer in the supine than in the other positions, and the unrestricted head movements are certainly less in that position (Lord 1948). It does not necessarily follow, however, that results obtained with the subject lying down will also hold for the sitting position.

It is obviously desirable that the results of eye movement measurements should be readily interpreted. As indicated in § 1, the chief item of interest is probably the movement of the retina relative to the image focused on the retina, although the movements of the eyeball itself are also of importance, since these are associated with changes in the tension of the extra-ocular muscles and, in all probability, with space perception. From a knowledge of the geometry of the recording apparatus, information about the movements of the eyeball can be immediately deduced from the results of methods (iii), (iv), (v) and (vi) of § 4. On the other hand, the subjective methods yield direct information about the movement of the image over the fundus, but not about the position of the eyeball. To obtain any other information from the various results certain assumptions have to be made. Thus in order to deduce the movement of the image relative to the retina by methods (ii), (iv), (v) and (vi), it has to be assumed that the centre of rotation of the eye is in a fixed and known position relative to the eye. In the case of method (vi) a value for the distance of the centre of rotation of the eye from the centre of curvature of the cornea has also to be assumed. The interpretation of the results of method (vii) depends, as already mentioned in § 4 (vii), on comparison with other methods.

Experiments referred to in § 4 (iii) suggest that the centre of rotation of the eye is not in fact fixed relative to the eye; hence, to the extent that the centre of rotation does move, so will the interpretation of eye movement records be uncertain.

The displacement of the centre according to Park and Park's results is not in practice very great and the error in regard to large rotations of the eye is not likely to be significant. It is just possible, however, that with small rotations of, say, less than 1°, the effective centre of rotation might be displaced very considerably and might even lie right outside the eye. This could occur if the rotation necessary to change the line of regard from one fixation spot to another were accompanied by some lateral translation. Suppose, for example, that when the eye turned to the right a sideways movement of 0·02 mm. to the right also occurred, owing, perhaps, to lack of proper co-ordination of the extra-ocular muscles. If the angle of rotation were only 6 minutes of arc, then a point about 12 mm. behind the anatomical centre of the eye, that is, a point very near the fovea, would have remained fixed in space and would therefore be the effective centre of rotation. For larger angles of rotation, the effective centre would gradually approach the anatomical centre. This is a purely hypothetical example, but it serves to show that some caution is necessary in the quantitative interpretation of any of the ordinary methods of recording small eyeball rotations. An effect of this kind might, for instance, be the explanation of the difference between the vertical and horizontal rotations reported by Berlin (1871), by Shen (1927) and by Lord and Wright (1949).

A further uncertainty in the interpretation of the records, but one which only affects deductions about movements of the image across the retina, has been raised in a paper by Park (1936 b). He has suggested that the variation in angle which he recorded with his anglometer, and also the deviation of the visual axes from accurate binocular convergence on a fixation target, as reported for example by Peckham (1934) and Clark (1936), is due to displacement or distortion of the crystalline lens, rather than to the use of some parafoveal part of the retina for fixation. His theory implies that fixation is a combined function of the extra-ocular muscles controlling the eyeball and the ciliary muscle which controls the shape of the crystalline lens. Thus he writes (Park 1936 b, p. 973): "Probably the extrinsic muscles rotate the eye to approximately the point of fixation and the ciliary muscle acts as what might be termed a fine adjustment for exact fixation by means of this change in the shape of the lens. In prolonged fixation it is quite possible that, due to innervation and contraction, the extrinsic muscles act to shift the fovea away from the image and that the ciliary muscle acts to shift the image to the new foveal position."

Since the ciliary muscle acts around the circumference of the crystalline lens, it would not seem to be well adapted to produce the unsymmetrical distortion or displacement of the lens that would be required to affect the direction of fixation, a displacement that would have to be fairly considerable in view of the relatively small difference in refractive index between the lens and the vitreous and aqueous humours on either side of it. However, Hess (1929) has reported an appreciable downward movement of the lens (0·25–0·36 mm.) during accommodation, and oscillation of the lens when the head is moved. Hess also observed irregularities on the lens border which he attributed to greater pressure on one side, and found that eserine removed the irregularities. These phenomena disappeared when the head was horizontal. Park has also claimed that both the behaviour of aphakic subjects and the effect of paralysing the ciliary muscle with drugs is consistent with his theory.

To sum up the position with regard to the eye movements themselves, nearly all the techniques indicate that the eyeball undergoes some movement even during

supposed steady fixation, and although none of the methods is entirely free from criticism and uncertainty, the agreement is rather impressive. The magnitude of this random movement has been put by some workers at a figure as high as 1 or 1½° rotation, e.g. by Judd, McAllister and Steele (1905) and by Clark (1936); at perhaps 20 or 30 minutes under average fixation conditions and 5 or 10 minutes under good fixation conditions by Lord and Wright (1948), with periods lasting for about ½ second at a time when the eye appeared to be steady to within 1 minute of arc; at two minutes of arc by Adler and Fliegelman (1934) with larger superimposed excursions of as much as 30 minutes; and at several minutes of arc by Riggs and Ratliff (1949). Only Hartridge and Thomson (1948) found no movement at all within the limits of sensitivity of their apparatus.

Apart from Park, it has generally been assumed that these eyeball movements have been accompanied by corresponding movements of the image across the retina. Thus, Dodge (1907) wrote that there is "no true fixation point", and he proposed the term 'fixation field'. This conception has been widely accepted by other workers and is supported by subjective experiments with after-images and counting techniques, where the interpretation is not affected by Park's theory of crystalline lens distortion. Subjective observations generally indicate a fixation field of the order of 5 or 10 minutes of arc, and this would appear to be a reasonable value to accept on present evidence, although Hartridge (1947) describes subjective experiments based on the assumption that very precise fixation could be maintained.

Before assessing the validity of the suggestion either of Peckham (1934) that the two foveas may fixate on points separated in the visual field by several degrees or of Park (1936 b) that distortion of the crystalline lens may displace the visual axis by several degrees, further experimental evidence is probably required.

The results of studies of the eyes in motion are on the whole in fair agreement. The saccadic nature of most eye movements, and especially those in reading, is firmly established, and also the smooth pursuit movement when following a moving target. A more complete study of the characteristic patterns of individual eye movements is desirable, especially in relation to various eye defects and to ability to perform different visual tasks, and some method of locating the centre of rotation of the eye during rotation would lead to a more conclusive interpretation of some of the records.

ACKNOWLEDGMENTS

The investigation of eye movements being carried out at the Imperial College is supported by the Medical Research Council, and the authors wish to express their thanks for this assistance. Several references have been made in this Report to *Reading and Visual Fatigue by* Carmichael and Dearborn; this book includes an excellent historical survey of the development of techniques for recording eye movements and has proved of material assistance in preparing the present, more condensed, summary.

REFERENCES

ADLER, F. H., and FLIEGELMAN, M., 1934, *Arch. Ophthal.*, **12**, 475.

BAUDOUIN, A., FISCHGOLD, H., CAUSSÉ, R., and LERIQUE, J., 1939, *Bull. Acad. Med. Paris*, **121**, 688.

BERLIN, E., 1871, *Arch. f. Ophthal.*, **17**, 193.

BRODAL, A., 1948, *Neurological Anatomy in relation to Clinical Medicine* (Oxford: University Press).

CARMICHAEL, L., and DEARBORN, W. F., 1948, *Reading and Visual Fatigue* (London: Harrap).
CLARK, B. C., 1934, *Amer. J. Psychol.*, **46,** 325; 1936, *Ibid.*, **48,** 82.
COBB, P. W., and MOSS, F. K., 1925, *J. Franklin Inst.*, **200,** 239.
CORDS, R., 1927, *Graefe's Arch. Ophthal.*, **118,** 771.
DELABARRE, E. B., 1898, *Amer. J. Psychol.*, **9,** 572.
DODGE, R., 1907, *Psychol. Monogr.*, **8,** No. 4; 1921, *J. Exp. Psychol.*, **4,** 165.
DODGE, R., and CLINE, T. S., 1901, *Psychol. Rev.*, **8,** 145.
DOHLMAN, G., 1925, *Acta Oto-laryng.*, Supp. **5,** 78; 1935, *Ibid.*, **23,** 50.
DUKE-ELDER, W. S., 1932, *Textbook of Ophthalmology*, Vol. I (London: Kimpton).
FULTON, J. F., 1943, *Physiology of the Nervous System* (Oxford: University Press).
GEORGE, E. J., TOREN, J. A., and LOWELL, J. W., 1923, *Amer. J. Ophthal.*, **6,** 833.
GRANIT, R., 1947, *Sensory Mechanisms of the Retina* (Oxford: University Press).
HARTRIDGE, H., 1947, *Phil. Trans. Roy. Soc.* B, **232,** 592.
HARTRIDGE, H., and THOMSON, L. C., 1948, *Brit. J. Ophthal.*, **32,** 581.
HELMHOLTZ, H., 1925, *Physiological Optics*, Vol. III, translated by J. P. C. SOUTHALL (*Opt. Soc. Amer.*).
HESS, C., 1929, *Handb. d. Norm. u. Path. Physiol.*, **12,** 145.
HOLMES, G., 1919, *Brit. Med. J.*, **2,** 193.
HUEY, E. B., 1898, *Amer. J. Psychol.*, **9,** 575; 1900, *Ibid.*, **11,** 283.
JACOBSON, E., 1930 a, *Amer. J. Physiol.*, **91,** 567; 1930 b, *Ibid.*, **95,** 694.
JAVAL, L. E., 1879, *Ann. d'Oculistique*, **82,** 242.
JONES, L. A., and HIGGINS, G. C., 1947, *J. Opt. Soc. Amer.*, **37,** 217.
JUDD, C. H., MCALLISTER, C. N., and STEELE, W. M., 1905, *Psychol. Monogr.*, **7,** No. 1.
KARSLAKE, J. S., 1940, *J. Appl. Psychol.*, **24,** 417.
LAMANSKY, S., 1869, *Pflüger's Arch.*, **2,** 418.
LANCASTER, W. B., 1941, *Amer. J. Ophthal.*, **24,** 485.
LANDOLT, A., 1891, *Arch. d'Ophthalm.*, **11,** 385.
LORD, M. P., 1948, *Proc. Phys. Soc.*, **61,** 489.
LORD, M. P., and WRIGHT, W. D., 1948, *Nature, Lond.*, **162,** 25; 1949, *Ibid.*, **163,** 803.
LORING, M. W., 1915, *Psychol. Rev.*, **22,** 354.
MARSHALL, W. H., and TALBOT, S. A., 1942, *Biological Symposia*, Vol. III, *Visual Mechanism* (Lancaster, U.S.A.: Cattell Press).
MARX, E., and TRENDELENBURG, W., 1911, *Zeit. f. Sinnesphysiol.*, **45,** 87.
MEYERS, I. L., 1929, *Arch. Neur. Psychiat.*, **21,** 901.
MILES, W. R., 1939 a, *Science*, **90,** 404; 1939 b, *Yale J. Biol. Med.*, **12,** 161; 1939 c, *Psychol. Bull.*, **36,** 536; 1940, *Science*, **91,** 456.
MOWRER, O. H., RUCH, T. C., and MILLER, N. E., 1936, *Amer. J. Physiol.*, **114,** 423.
NEWHALL, S. M., 1928, *Amer. J. Psychol.*, **40,** 628.
O'BRIEN, B., and DICKERMAN, E., 1948, *J. Opt. Soc. Amer.*, **38,** 1096.
OGLE, K. N., MUSSEY, F., and DE H. PRANGEN, A., 1949, *Amer. J. Ophthal.*, **32,** 1069.
OHM, J., 1914, *Z. Augenheilk.* **32,** 4; 1916, *Ibid.*, **36,** 198; 1928, *Graefe's Arch. f. Ophthal.*, **120,** 235.
ORCHANSKY, J., 1899, *Zentralbl. f. Physiol.*, **12,** 785.
PARK, G. E., 1936 a, *Arch. Ophthal.*, **15,** 703; 1936 b, *Amer. J. Ophthal.*, **19,** 967.
PARK, G. E., and PARK, R. S., 1933, *Amer. J. Physiol.*, **104,** 545; 1940, *Arch. Ophthal.*, **23,** 1216.
PARSONS, J. H., 1949, *Lancet*, **256,** 1120.
PECKHAM, R. H., 1934, *Arch. Ophthal.*, **12,** 562.
POLYAK, S. L., 1941, *The Retina* (Chicago: University Press).
RIGGS, L. A., and RATLIFF, F., 1949, *J. Opt. Soc. Amer.*, **39,** 630.
RODIN, F. H., and NEWELL, R. R., 1934, *Arch. Ophthal.*, **12,** 525.
SCHOTT, E., 1922, *Deutsch. Arch. f. klin. Med.*, **140,** 79.
SHEN, E., 1927, *J. Exp. Psychol.*, **10,** 158.
STRATTON, G. M., 1902, *Phil. Stud.*, **20,** 336; 1906, *Psychol. Rev.*, **13,** 82.
TAYLOR, E. A., 1937, *Controlled Reading* (Chicago: University Press).
TINKER, M. A., 1931, *Amer. J. Psychol.*, **43,** 115; 1946, *Psychol. Bull.*, **43,** 93.
WALLS, G. L., 1942, *The Vertebrate Eye* (Michigan: Cranbrook Institute of Science).
WEAVER, H. E., 1931, *Psychol. Bull.*, **28,** 211.
WOLFF, E., 1948, *The Anatomy of the Eye and Orbit* (London: Lewis).

RECENT ADVANCES IN INFRA-RED SOLAR SPECTROSCOPY

By L. GOLDBERG

McMath–Hulbert Observatory of the University of Michigan, Pontiac, Michigan

CONTENTS

	PAGE
§ 1. Introduction	24
§ 2. Instrumentation	26
§ 3. An atlas of the infra-red solar spectrum	28
§ 4. Solar lines	29
Factors governing the appearance of solar lines	34
§ 5. Telluric lines	35
(i) Carbon dioxide	36
(ii) Methane	41
(iii) Nitrous oxide	42
(iv) Water vapour	44
Acknowledgments	44
References	44

§ 1. INTRODUCTION

The existence of both infra-red and ultra-violet radiation was demonstrated by Sir William Herschel (1800), who observed the heating effects of invisible solar rays on a thermometer placed in the focal plane of his spectroscope. The first 'map' of the infra-red solar spectrum was Sir John Herschel's (1840) Thermograph, which he prepared by projecting the spectrum on a strip of dull black paper moistened with alcohol. The damp spots remaining on the paper after exposure to radiation indicated the presence and location of absorption bands. The primitive detection methods employed by the Herschels were soon superseded by the advent of the thermocouple which was discovered by Seebeck (1826), followed by Nobili's (1831) invention of the thermopile, and the application of the phosphor method by A. E. Becquerel (1843 *et seq.*). The latter eventually mapped the spectrum to an infra-red limit of about $1{\cdot}4\,\mu$. Draper (1843) also used the phosphor technique to observe three wide infra-red bands (alpha, beta, and gamma), the existence of which were later verified by Foucault and Fizeau (1847) with thermocouples. The phosphor method of observation was considerably improved by A. H. Becquerel (1883), who mapped the solar spectrum to $1{\cdot}4\,\mu$, with a CS_2 prism and Rutherford grating. He was also the first to note the changing appearance of telluric absorption bands due to variations in atmospheric conditions and in solar altitude.

The years following these early developments have seen steady progress in the exploration of the infra-red solar spectrum, due both to the design of more efficient optical systems, and to the development of highly sensitive energy detectors. Advances in detection methods have resulted from improvements both in photographic emulsions and in devices for more or less instantaneous energy recording. The first application of the photographic method was made by Draper (1881) with a silver bromide plate placed on a phosphorescent screen

exposed to infra-red radiation. Abney (1880) employed what is now conventional photography in recording wavelengths slightly longer than 1 μ, and later (1886) achieved the first photographic map of the infra-red spectrum, from 0·69 μ to 1 μ, using a Rowland concave grating. He was the first to resolve and measure infra-red absorption lines, claiming a precision of 0·1 A. for his wavelength determinations. Improved photographic maps of the solar spectrum extending to about 1 μ were published by Higgs (1898), Meggers (1918), Burns (1920), and Brackett (1921).

More recently, St. John, Moore, Ware, Adams, and Babcock (1928) have listed 1,700 wavelengths between λ7330 and λ10,216 in a supplement to the *Revision of Rowland's Preliminary Table of Solar Spectrum Wavelengths*. Finally, the application of the photographic process to the infra-red solar spectrum has culminated in the publication by Babcock and Moore (1947), after 22 years of preparation, of an infra-red extension to the Revised Rowland. The 'IRE' contains wavelengths for about 7,400 lines in the region λ6600–λ13,495, 51% of which are due at least in part to atmospheric absorption. Of the remaining lines, 38% are certainly or probably solar, and 11% of the lines are so weak that their origin is uncertain. Well over one-half of the solar lines have been identified with known atoms and molecules. The precision of the measured wavelengths is extremely high, the errors being less than one part in a million for wavelengths shorter than 1·21 μ.

Although the photographic method has yielded much higher spectral resolution than have other types of infra-red detection, its application thus far has been limited to wavelengths shorter than 1·4 μ. On the other hand, bolometers, thermocouples and thermopiles have been successfully employed to map the solar spectrum to an optical limit of 24 μ. Lamansky (1872) used an insensitive thermopile to map the solar prismatic spectrum, recording three large gaps in the infra-red spectrum in verification of Draper's observations. Lamansky was the first to attribute these bands to absorption by the earth's atmosphere. In 1881, Langley began his epic investigations of the infra-red solar spectrum, which were later continued at the Smithsonian Institution by Abbot, Fowle, and Freeman. Langley's achievements were made possible by his independent invention and introduction to infra-red spectroscopy of the sensitive bolometer. The earlier discovery of this detector by Svanberg (1857) appears to have escaped general notice. By 1888, Langley had obtained a continuous record of the solar spectrum from the ultra-violet cut off at 0·3 μ to a long wavelength limit of 5·3 μ. Langley was able to detect, but not to record, radiation of longer wavelengths. Fowle (1917) extended the spectrum to 13 μ, and Abbot and Freeman (1929) mapped the solar spectrum from λ7582 to λ18,194 with sufficient accuracy to resolve strong solar absorption lines near 1 μ.

The thermocouple has been used with great success by Adel and by Migeotte. The former (1942) has recently extended the prismatic solar spectrum from 14 μ to 24 μ and has mapped (1941a) the grating spectrum from 2 μ to 14 μ with a theoretical resolving power of about 10,000. Migeotte (1945, 1948) has achieved the highest resolution thus far obtained with grating and thermocouple in the regions between 1·3 μ and 1·5 μ and between 2·7 μ and 5·5 μ.

As at the end of 1946, the status of infra-red solar spectroscopy could be roughly summarized by the statement that the spectrum had been mapped photographically with fine detail to about 1·4 μ and with considerably lower

resolution with various types of heat detectors to 24 μ. (We do not consider here the post-war developments in radio-wave spectroscopy, by means of which solar radiation has been detected in the wavelength region 1 cm. to 15 m.) The most recent advance in infra-red solar spectroscopy in the region 1·4–3·6 μ has been achieved at the McMath–Hulbert Observatory by a group of workers headed by Dr. Robert R. McMath, and including Dr. Orren C. Mohler, Dr. A. Keith Pierce and the writer. The Michigan group have made use of lead sulphide photoconductive cells and a high-dispersion grating spectrometer to map the spectrum with resolution comparable with that which can be obtained photographically at shorter wavelengths.

Lead sulphide cells were first employed as infra-red detectors in Germany during World War II and were later developed in Great Britain by Starkiewicz, Sosnowski, and Simpson (1946) at the Admiralty Research Laboratory and in the United States by Cashman (1946) at Northwestern University. The PbS cells are sensitive from well below 1 μ to 3·6 μ. Maximum sensitivity can occur at any wavelength between 1 μ and 3 μ depending on the mode of construction of the photoconductive layer. The cells operate satisfactorily at room temperature; but refrigeration to −80° C. increases the signal-to-noise ratio by a factor of about ten and also shifts the sensitivity maximum to longer wavelengths. The response time of the cells is of the order of microseconds and the resistance is in the neighbourhood of one megohm. These characteristics lend themselves to A.C. amplification, with interruption frequencies of the order of 1,000 c/s. Amplifiers designed for use with the Cashman PbS cells have been developed by Wilson (Kuiper, Wilson, and Cashman 1947).

The performance of the Admiralty Research Laboratory cells at a temperature of −78° C. was studied by Sutherland, Blackwell and Fellgett (1946), who found a signal-to-noise ratio at 2·6 μ about 100 times that of the Hilger–Schwarz thermocouple. With a non-echelette grating of 14,400 lines per inch and slit widths of 0·08 cm^{-1} they were able to resolve lines of H_2O at 3,970 cm^{-1} as close as 0·14 cm^{-1} apart. Previously, the closest pair of lines resolved in the same band had a separation of 0·6 cm^{-1}. Sutherland, Blackwell and Fellgett concluded that an additional gain in resolving power, up to a factor of two, could be attained with improved spectrometer optics.

The majority of the PbS cells constructed by Cashman have been used at room temperature. Cashman states that the peak sensitivity of his uncooled cells is about 1,000 times greater than the sensitivity of thermocouple receivers.

§ 2. INSTRUMENTATION

The instrumentation involved in the application of the Cashman cells to the infra-red solar spectrum has been described in detail by McMath and Mohler (1949). The schematic diagram in Figure 1 shows the optical layout of the all-mirror image-forming and dispersing systems in the 70-foot McGregor solar tower. The four mirrors and CaF_2 prism between S_1 and S_2 form a predispersing unit. The Pfund-type spectrometer has a focal ratio of 50, the focal lengths of the collimating and focusing mirrors being 277 inches. Scanning of the spectrum by the PbS cell is accomplished by rotation of the grating at any one of three rates: $\frac{3}{10}$, $\frac{3}{4}$ or $1\frac{1}{2}$ degrees per hour. Two gratings, ruled 15,000 lines per inch and 5,000 lines per inch, have been employed for the spectral regions 1·0–2·5 μ and 2·9–3·6 μ, respectively.

The PbS cells in use at the McMath–Hulbert Observatory have all been especially constructed by Dr. Cashman, who has provided cells of both uncooled and cooled types with envelopes of glass and of fused silica respectively. The silica cell is shaped like a Dewar flask and is cooled by a mixture of solid carbon dioxide and acetone. The A.C. amplifier was designed and constructed by Wilson to match the characteristics of the Cashman cells.

The output of the amplifier is passed to a Leeds and Northrup Speedomax Recorder, which traces the spectrum in Indian ink on unruled white paper. For the three different rates of grating rotation, the dispersions of the spectrum

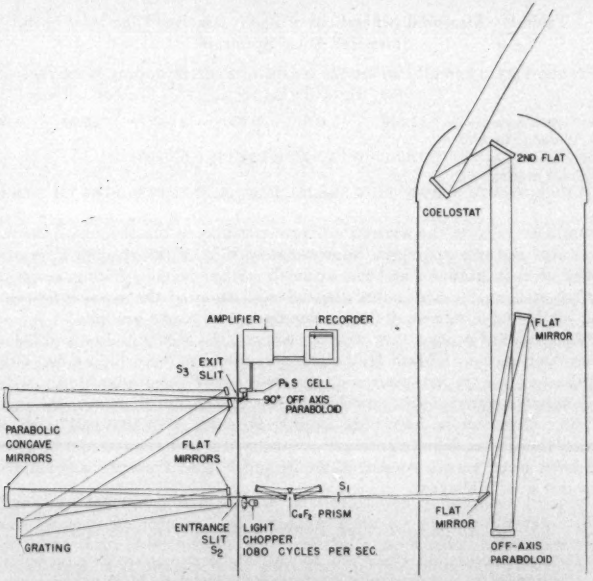


Figure 1. Schematic view of the McGregor tower and spectrometer of the McMath–Hulbert Observatory.

in mm/A. at λ20,000 are 3, 6, and 12 respectively. A careful study has been made of the relation between relative radiation intensity and recorded deflection, with the result that the response has been found to be very closely linear for all deflections greater than 2 cm. The amplifier gain is usually adjusted so that the continuous spectrum produces a deflection of about 20 cm. This means that intensity measurements may be made directly on the tracings for all but the very strongest lines.

Infra-red wavelength standards have been established by Rowland's method of overlapping orders with reference to Fraunhofer lines visible in the third and fourth order grating spectra. It is estimated that an accuracy of about one

part in 200,000 is obtainable by this method, the limiting factors being uncertainty as to the index of refraction of air in the infra-red, and slight irregularities in the motions of the grating and recording paper. The accuracy obtained is more than adequate for purposes of identification and is sufficient for the fairly accurate evaluation of molecular constants.

A measure of the overall performance of the spectrometer is afforded by the separation of close pairs of lines just clearly resolved on the tracings. Such measures are given in Table 1 for various wavelengths from λ 12,000 to λ 36,000 and for both the Mount Wilson and Michigan gratings. The numbers in

Table 1. Measured Separations of Barely Resolved Line Pairs in the Infra-Red Solar Spectrum.

(The numbers in parentheses are the separations corresponding to the theoretical resolving power.)

Wavelength region (A.)	12,000	16,000	20,000	24,000	32,000	36,000
Mt. Wilson grating, 66,700 lines (cm^{-1})	0·23(0·12)	0·12(0·09)	0·10(0·07)	0·07(0·06)	—	—
Michigan grating, 24,139 lines (cm^{-1})	0·52(0·35)	0·33(0·26)	0·28(0·21)	0·23(0·17)	0·14(0·13)	0·13(0·11)

parentheses refer to the separations corresponding to the theoretical resolving powers of the two gratings. Approximately 80% of the theoretical resolving power of each grating has been attained. Since neither grating is optically perfect, it seems probable that the full realization of the possibilities of the lead-sulphide cell must await the development of improved gratings.

As a result of a cooperative arrangement with the Mount Wilson and Palomar Observatories, the McMath–Hulbert Observatory has recently installed a near-duplicate of its infra-red spectrometer on the Snow Telescope on Mount Wilson. The objective mirror of the Snow Telescope is 24 inches in diameter and is figured at f/30. Observations have been in progress since May 1949. The tracings secured thus far are definitely of superior quality, due to the increased transparency and lower water vapour content of the Mount Wilson atmosphere as compared with that over Michigan.

§3. AN ATLAS OF THE INFRA-RED SOLAR SPECTRUM

The instrumentation described above has been employed in a series of investigations of the infra-red solar spectrum at the McMath–Hulbert Observatory. Tracings have been secured both at Lake Angelus and at Mount Wilson covering the entire spectral region from 0·8 to 3·6 μ. The Mount Wilson records have been incorporated into an atlas of the infra-red solar spectrum, to be published shortly. The atlas will consist of approximately 250 tracings, each about 70 A. in length, and embracing the spectral region from 0·8 to 2·5 μ. In order to achieve some degree of uniformity in the intensities of the telluric lines, the atlas observations were confined to the period from two hours before to two hours after noon on days of relatively low humidity. The infra-red atlas is intended to overlap and to extend the elaborate Utrecht Atlas (Minnaert, Mulders and Houtgast 1940) which terminates at λ 8771.

Certain of the more interesting records are shown in Figures 3 to 15. Due to selective absorption by the earth's atmosphere, several broad areas of the

infra-red solar spectrum are entirely masked, but three relatively transparent 'windows' are available for observation of solar and telluric lines. The location of the 'windows' is well shown in Langley's (1900) early map of the spectrum which is reproduced in Figure 2. The absorptions at 1·4 μ and at 1·8 μ are produced by water vapour, whereas the absorption at 2·6 μ is due both to water vapour and to carbon dioxide. The three transparent regions are centred approximately at 1·65 μ, at 2·20 μ and at 3·3 μ. The absorption lines in the clear regions are divided between those that are formed in the earth's atmosphere and those that originate in the sun. The solar lines are easily distinguished from comparisons between tracings made with the sun at high and at low

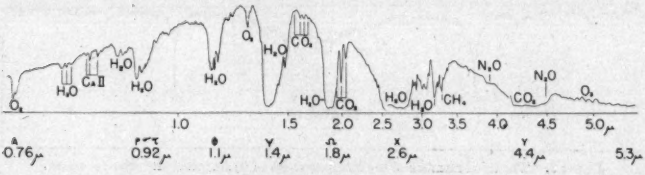


Figure 2. The solar spectrum as observed by S. P. Langley with the assistance of C. G. Abbot, 1898.

altitudes. Well over 400 solar lines have been newly discovered in this way, approximately 75% of them in the 1·65 μ region of the spectrum. Although in the 1·65 μ region the ratio of solar to telluric lines on noonday tracings is approximately 50%, in the 2·20 μ region the ratio is about 12%, and at 3·3 μ less than half a dozen solar lines have been found. The progressive reduction in numbers of observed solar lines with increasing wavelength may be ascribed in part to increased telluric absorption at the longer wavelengths, but more important reasons, relating to the structure of atoms and to the physics of the solar atmosphere, will be discussed later in this paper.

§4. SOLAR LINES

Published laboratory data on the wavelengths of atomic lines in spectral regions beyond the infra-red photographic limit are exceedingly rare. The observations have been made with relatively insensitive bolometers and thermocouples, and usually with unsteady arcs as sources. Consequently, only the very strongest lines have been recorded and the measured wavelengths suffer from uncertainties as large as 2 A. R. A. Fisher of Northwestern University (private communication) has recently developed a steady source for the Fe I spectrum and is engaged in the mapping and analysis of the spectrum from 1·0 μ to 3·0 μ, also employing a Cashman cell, Wilson amplifier, and Pfund-type spectrometer optics. Similar spectrometer and detector installations have been completed at the National Bureau of Standards by Humphreys and Plyler (1949) and at the Ohio State University by Nielsen and his co-workers (1949).

Pending the availability of accurate laboratory wavelengths, the McMath–Hulbert workers (Goldberg, Mohler and McMath 1949) have had recourse to the prediction of wavelengths from the tables of atomic energy levels compiled by Moore (1949). By this method, approximately 160 infra-red solar lines

have been identified, or over one-third of the total found between 1·5 and 3·6 μ. The identified elements, all of which are known to be abundant in the solar atmosphere, are H, C, Na, Mg, Al, Si, Ca, Fe, Mn and Ni. Only lines of neutral atoms have thus far been found in the infra-red spectrum.

The number of absorption lines identified for each element is given in Table 2.

Table 2. Elements present in the Infra-Red Solar Spectrum between 1·5μ and 3·6μ

Element	H	C	Na	Mg	Al	Si	Ca	Fe	Mn	Ni
No. of lines	3	9	4	7	5	48	11	64	2	8

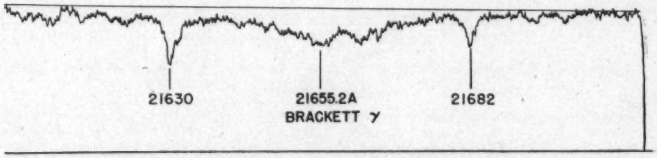


Figure 3. The third member of the Brackett series of hydrogen in the solar spectrum.

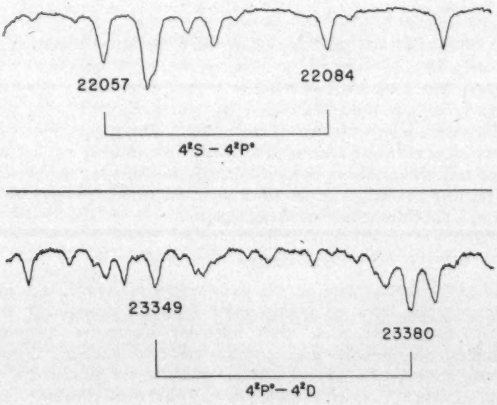


Figure 4. Na I lines in the infra-red solar spectrum. The horizontal line at the bottom of each tracing in this and succeeding figures represents zero intensity.

The three lines of hydrogen are the third, seventh and eighth members of the Brackett series at λλ 16,407·2, 16,806·6 and 21,655·4. The third member of the series is shown in Figure 3. The enormous width of the line, about 40 A., is characteristic of hydrogen atoms subject to statistical Stark broadening, particularly when energy levels of high quantum numbers are involved. The

widths of the Brackett series lines are also unusually great because of the fact that for given quantum numbers and field strengths the Stark widths are proportional to the square of the wavelength. The two higher members of the Brackett series are also extremely wide, and so shallow that they can barely be distinguished from the continuous spectrum.

The Na I doublet corresponding to the transition 4s–4p at λλ 22,083 and 22,056 is shown in Figure 4. The doublet 4p–4d also appears prominently.

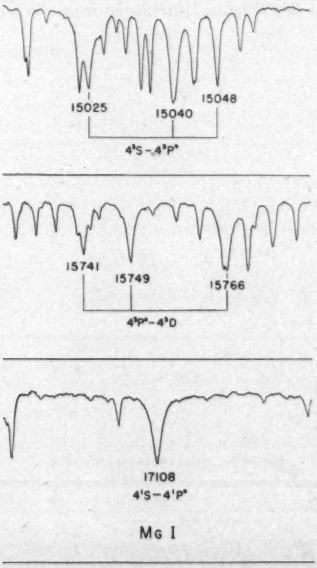


Figure 5. Neutral magnesium in the infra-red solar spectrum. Neighbouring lines are due to H_2O at λ 15,000 and to CO_2 at λ 15,700.

The lines of Mg I are very wide and intense in the infra-red solar spectrum. The $4s\,^1S–4p\,^1P^0$ line at λ 17,108·5 is the strongest solar infra-red line observed, with the exception of the third member of the Brackett series of hydrogen. Other interesting lines of Mg I are included in the multiplets $4p\,^3P^0–4d\,^3D$ at λ 15,750 and $4s\,^3S–4p\,^3P^0$ at λ 15,040 (see Figure 5). Laboratory observations have disclosed only two members of each of the latter multiplets, the implication being that the $^3P^0$ levels $J=0$ and $J=1$ were not resolved. Evidence has been given (Goldberg, Mohler and McMath 1949), based primarily on agreement with the Landé interval rule, for the complete resolution of the $4\,^3P^0$ term in the solar spectrum.

Al I is represented in the infra-red solar spectrum by the three lines of the multiplet $4p\,^2P^0–4d\,^2D$ at approximately λ 16,700 and by the two members of

the multiplet $4p\,^2P^0$–$5s\,^2S$ at λ21,100 (see Figure 6). The Si I lines (Figure 7) are among the strongest and most numerous in the entire infra-red spectrum. They occur in the regions of all three transparent windows including three lines at approximately 3·3 μ.

The majority of the lines of Ca I (Figure 8) appear near 2·0 μ where they suffer from blending with the strong telluric lines of CO_2. As in the photographic region of the spectrum, Fe I is the most abundant source of lines in the near infra-red, approximately 64 lines having been so identified. Only three or four multiplets of Fe I (Figure 7) predicted from the term values of Russell,

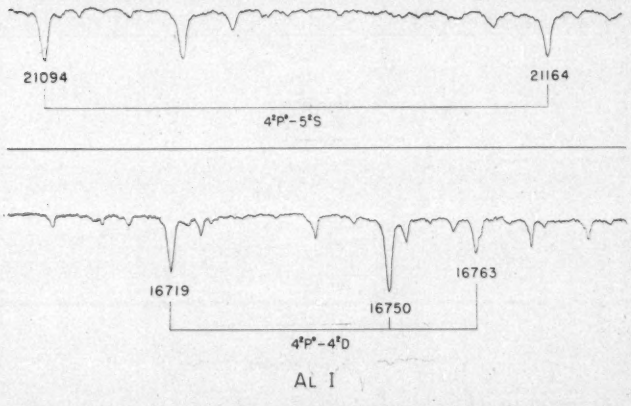


Figure 6. Solar infra-red lines of Al I.

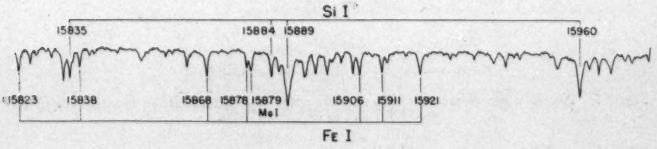


Figure 7. Solar infra-red lines chiefly of Fe I and Si I.

Moore and Weeks (1944) are entirely normal in the sense that they obey all of the rules of LS coupling. The laboratory spectrum of Fe I in the photographic region shows, however, that virtually any type of transition may be expected as long as it obeys the Laporte parity rule. In the absence of laboratory data and of theoretical intensities for the multiplet lines, however, the identifications of 'abnormal' multiplets are highly uncertain. It is estimated that as many as 100 solar lines, unidentified at present, will be identified with Fe I after the infra-red laboratory data become available.

Additional lines of low ionization potential atoms, especially those of Ti I and V I, may be expected to appear in the infra-red sunspot spectrum. The Ti I multiplet $a\,^3G$–$z\,^3G^0$ appears to be present near λ15,500 in the spectrum of a small sunspot recorded during the summer of 1948 (McMath and Goldberg 1949). A short section of the sunspot spectrum near λ15,900 is shown in Figure 9, where it is compared with the disc spectrum at the same wavelengths.

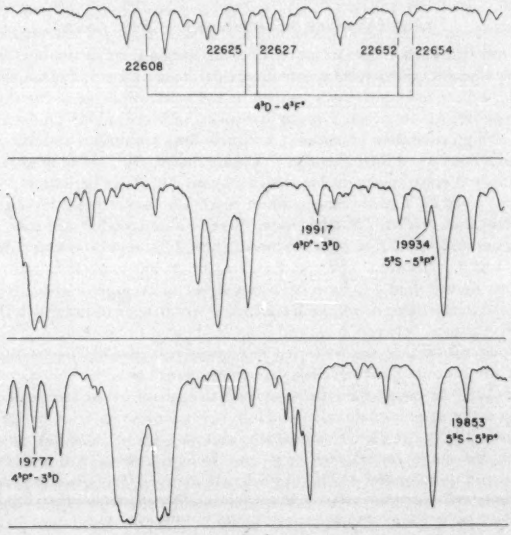


Figure 8. Solar infra-red lines of Ca I.

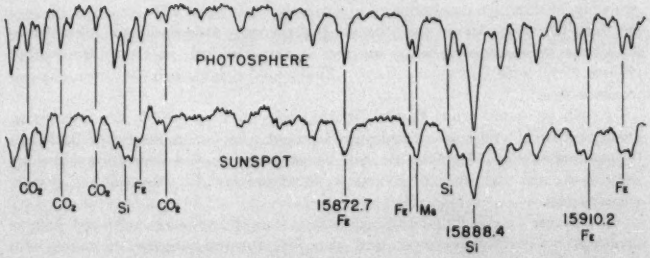


Figure 9. A section of the infra-red spectrum of a sunspot as compared with that of the neighbouring disc.

It will be noted that whereas the telluric lines of CO_2 are alike in appearance in both spectra, the solar lines of Fe, Si, and Mg are extremely wide in the spot spectrum. The widening is undoubtedly a result of the Zeeman effect of the sunspot's magnetic field. The infra-red spectrum offers interesting possibilities for the study of magnetic effects in the solar atmosphere in view of the fact that Zeeman splitting in wavelength units is proportional to the square of the wavelength.

Factors Governing the Appearance of Solar Lines

The most important factors governing the appearance of the spectral lines of a given element in the solar spectrum are (*a*) its abundance, (*b*) the excitation potential, and (*c*) the continuous opacity of the solar atmosphere for the wavelength considered. As pointed out by Babcock and Moore (1947), most infra-red lines are of high excitation potential. Ultimate lines are rare or absent, and even penultimate lines occur infrequently. Approximately 90% of all lines identified in the infra-red solar spectrum between $1{\cdot}5\,\mu$ and $3{\cdot}6\,\mu$ have excitation potentials in excess of 5 volts. The lowest excitation potential observed is 2·17 volts for an inter-system line of Fe I. Furthermore, there is a remarkable decrease in both the numbers and intensities of solar lines in the $2{\cdot}2\,\mu$ region as compared with those at $1{\cdot}65\,\mu$. Between $1{\cdot}52\,\mu$ and $1{\cdot}75\,\mu$ about 300 solar lines with central absorptions higher than 4% have been found, or an average of about 0·32 lines per cm^{-1}. On the other hand, the frequency of occurrence of lines with the same minimum intensity between $2{\cdot}1\,\mu$ and $2{\cdot}4\,\mu$ is about 0·11 per cm^{-1}. This result lends strong support to the idea that the negative hydrogen ion is the major source of continuous opacity in the solar atmosphere.

Wildt (1939) first suggested the negative hydrogen ion as the source of opacity. Following earlier approximate calculations by several investigators, Chandrasekhar and Breen (1946) calculated with high accuracy the theoretical absorption coefficient for both bound–free and free–free transitions Simultaneously, Chalonge and Kourganoff (1946) ingeniously derived the observed variation with wavelength of the solar absorption coefficient from limb darkening observations by Abbot, Plaskett and Fabry. The observed and calculated curves, shown in Figure 10, are seen to agree closely between $\lambda 4000$ and $\lambda 16{,}000$, and to justify the conclusion (Chandrasekhar and Münch 1946) that H^- is the principal source of solar opacity. The discrepancy at longer wavelengths is probably within the uncertainties of both the calculated absorption coefficient for free–free transitions and the limb-darkening observations. New limb-darkening measures are being secured at the McMath–Hulbert Observatory (Pierce 1949) with Cashman cells to a limit of $3{\cdot}6\,\mu$, and with the thermocouple to about $20\,\mu$.

It will be noted from Figure 10 that, beginning at $\lambda 4000$, the absorption coefficient of H^- rises to a maximum at about $\lambda 8600$ and thereafter decreases to a minimum at $\lambda 16{,}600$. At this point, the free–free transitions become important and the absorption curve increases steadily toward the longer wavelengths.

If we bear in mind that the production of high excitation infra-red lines is favoured by high temperature and also that the temperature increases with depth in the solar atmosphere, it is clear that the opacity minimum at $\lambda 16{,}600$

is highly favourable for the appearance of infra-red absorption lines. At longer wavelengths the increased opacity acts to screen the high temperature layers of the atmosphere, and on the average the infra-red lines will decrease both in numbers and in intensity.

The foregoing considerations are not in harmony with the tentative identification by Adel (1941 b) of five strong lines in the solar spectrum near 10 μ as high-level transitions in Na I and K I. Three lines, at 1,099·59 cm^{-1}, 1,101·33 cm^{-1}, and 1,091·04 cm^{-1}, are ascribed to the transition $3d\,^2D$–$4p\,^2P^0$ in Na I. Although no line intensities are given, the appearance of the lines on the published reproductions of the tracings suggests that their total absorptions are of the order of 0·30 cm^{-1}. On the other hand, the total absorptions of the lines of the 4s–4p

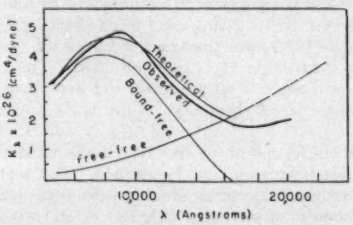


Figure 10. The absorption coefficient of the negative hydrogen ion per neutral H atom and per unit electron pressure at T=6,300° K., calculated by Chandrasekhar and Breen. The heavy curve is the observed coefficient by Chalonge and Kourganoff.

(*Reproduced from a review article by A. J. Deutsch, Rev. Mod. Phys.*, 1948, **20**, 388).

doublet at λ22,000 are approximately 0·050 cm^{-1}. The f-value for the 3d–4p transition has not been calculated, but in all probability is somewhat less than that for the 4s–4p transition. Furthermore, the 3d level lies about 0·4 volts above the 4s level and the calculated continuous absorption coefficient at 9·11 μ is about 40 times greater than at 2·21 μ. These circumstances are difficult to reconcile with the appearance of the 3d–4p multiplet in the solar spectrum. The theoretical argument against the identification of three solar lines at about 1,175 cm^{-1} with the $6\,^2P^0$–$5\,^2D$ transition in K I is even more convincing. The solar lines are strong almost to the point of saturation. Yet the transition $4\,^2P^0$–$3\,^2D$, for which the lower excitation potential is less than that of $6\,^2P^0$–$5\,^2D$ by nearly two volts, appears in the solar spectrum at λ11,769 with an extremely low Rowland intensity of −2 (Babcock and Moore 1947).

§5. TELLURIC LINES

More than one-half of the lines in the region between 1·4 μ and 3·6 μ in the solar spectrum arise from absorption by molecules in the earth's atmosphere. At 1·4 μ, 1·8 μ and 2·6 μ, absorption by H_2O and CO_2 renders the atmosphere completely opaque. In the intervening regions the telluric absorption is much weaker but produces well defined band systems of H_2O, CO_2, CH_4 and N_2O. The near infra-red band systems of these molecules have recently been investigated by McMath, Goldberg, Mohler and Pierce, as described below.

(i) *Carbon Dioxide*

The near infra-red absorption by molecules of carbon dioxide in the earth's atmosphere is plainly shown on early tracings of the solar spectrum. Langley (1900) recorded two bands at 2·01 μ and 2·05 μ, which he designated as ω_1 and ω_2. Laboratory observations by Barker and Wu (1934) later revealed a third band at 1·95 μ which is not apparent on Langley's tracing, due to strong absorption by water vapour. Adel (1943) later recorded all three bands with sufficient resolution to reveal their rotational structures. The band at 1·95 μ was designated ω_0 by Adel in conformity with Langley's notation. The 2 μ bands result from transitions between the ground level of CO_2 and the triad of excited vibrational levels 04_00, 12_01 and 20_01, the excited levels being in Fermi resonance. Barker and Wu also recorded in the laboratory four weaker unresolved bands of CO_2 at 1·54 μ, 1·57 μ, 1·60 μ and 1·65 μ, which were later identified by Adel and Lampland (1938) with transitions between the ground level and the excited resonating levels 06_01, 14_01, 22_01 and 30_01. The CO_2 bands at 1·6 μ are plainly visible on Langley's early tracings and were recognized by Adel and Lampland as a single small absorption in the low-resolution prismatic solar spectrum.

All seven bands of CO_2 appear fully resolved on the McMath–Hulbert tracings of the infra-red solar spectrum (see Figure 11). The high precision of the measured wave numbers has warranted an independent determination both of the equilibrium moment of inertia of CO_2 and of the rotational constants α_1, α_2 and α_3 (Goldberg, Mohler, McMath and Pierce 1949). The results are shown in Table 3, where they are compared with earlier determinations from

Table 3. Rotational Constants of CO_2 in cm^{-1}

	B_e	B_{000}	$B_e\alpha_1$	$B_e\alpha_2$	$B_e\alpha_3$
Lake Angelus	0·3920	0·3906	−0·00103	+0·000640	−0·00306
Dennison	0·3925	0·3911	0·00058	+0·00045	−0·00307
Herzberg	0·3906	0·3895 *	−0·00056	+0·00062	−0·0029

* Newly revised value 0·39043 cm^{-1}.

other infra-red bands by Dennison (1940) and Herzberg (1945). In the main, the agreement with Dennison's results is excellent. Herzberg (1948) has recently secured new results from a study of the photographic infra-red bands of CO_2 that agree quite closely with the McMath–Hulbert results. For example, his new value of B_{000} is 0·39043 as compared with the McMath–Hulbert value of 0·3906.

The great strength of the 2 μ bands of CO_2 suggests the presence in the spectrum of isotope bands resulting from similar transitions in the molecules $^{13}C^{16}O_2$ and $^{12}C^{16}O^{18}O$, the relative abundances of the isotopes ^{13}C and ^{18}O being 1·1% and 0·2% respectively. The presence of such bands on the laboratory tracings of Barker and Wu is suggested by a small dip in the continuous spectrum in the space between the CO_2 bands at 2·01 μ and 2·05 μ. The infra-red solar spectrum obtained at Lake Angelus has a series of faint but clearly resolved bands adjacent to and superposed on the bands of ordinary CO_2 (see Figure 12). A total of six isotope bands would be expected to appear at 2 μ, three each corresponding to the two isotopic forms of CO_2. As compared with the bands of ordinary CO_2 the isotope bands should be shifted to lower frequencies by varying amounts from 57 cm^{-1} to 115 cm^{-1}. Five of the six expected bands have been found (Goldberg, Mohler and McMath 1948) in very nearly the

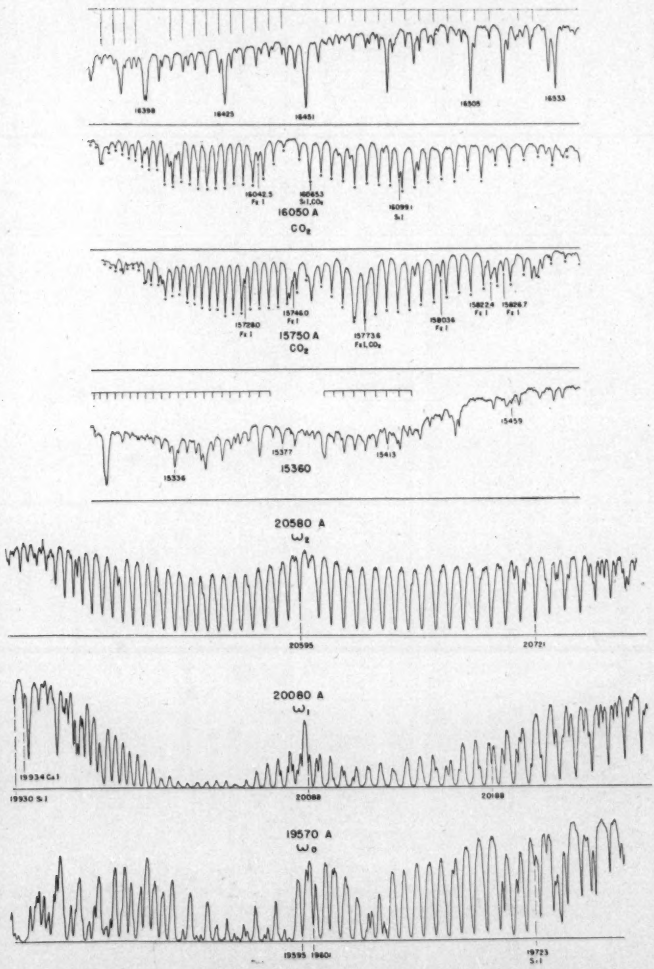


Figure 11. Telluric combination bands of CO_2 at 1·6 μ and at 2 μ. The lines of Fe I, Si I and Ca I are of solar origin. The regularly-spaced series of lines at λλ 16,398, 16,425, 16,451 etc., are members of the R-branch of the $2\nu_3$ band of CH_4 (see Figure 14).

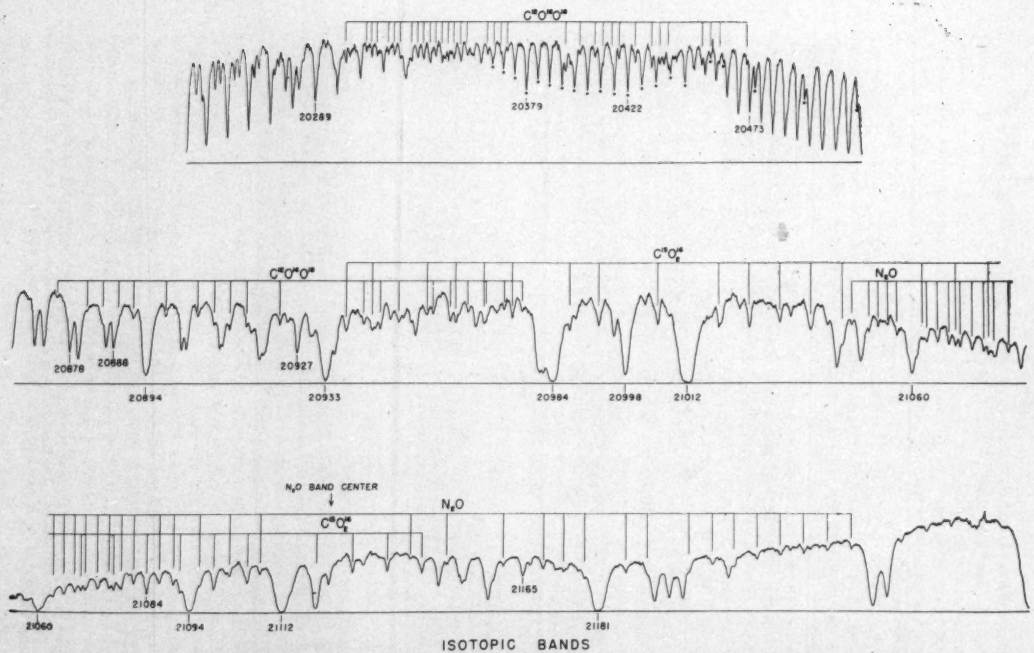


Figure 12. Telluric bands of $^{12}C^{16}O^{16}O$, $^{13}C^{16}O_2$ and of N_2O in the infra-red solar spectrum. The dots mark the lines of a band of $^{13}C^{16}O_2$. The dispersion of the upper tracing is about one-half that of the tracings below.

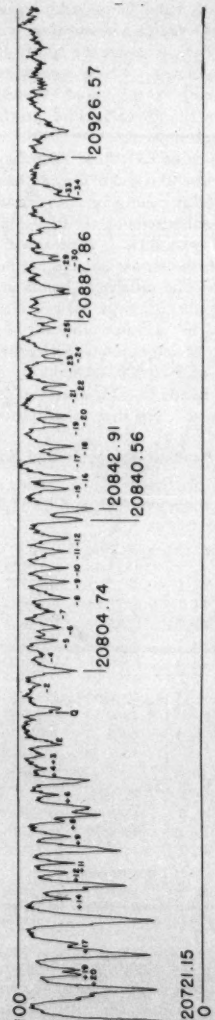


Figure 13. Difference band of CO_2 at 2·08 μ in the solar spectrum.

calculated positions, the sixth band being submerged by the strong band ω_1. Furthermore, the bands of $^{13}C^{16}O_2$ have identical rotational spacings, as compared with the main bands of $^{12}C^{16}O_2$, whereas the bands of $^{12}C^{16}O^{18}O$ have approximately one-half the usual spacing. These results are in accord with theory, since substitution of an atom of ^{18}O for one of ^{16}O destroys the symmetry of the CO_2 molecule and results in the appearance of alternate rotational lines which are absent for the symmetrical molecule.

Two additional new bands of CO_2 have been found in the solar spectrum (McMath and Mohler 1948) at 2·01 μ and 2·08 μ and have been identified (Mohler, McMath and Goldberg 1949) as belonging to the triad of resonating bands that originate from the excited vibrational level 01_10 and terminate in the levels 05_11, 13_11 and 21_11 (see Figure 13). The third member of the group at 5,131·6 cm^{-1} is blended with the strong band ω_0, although numerous individual lines appear to be present. The difference bands are interesting in that they display the *l*-type doubling characteristic of π–π transitions. Analysis of the rotational structures leads to a value for the *l*-type doubling constant $q = 0{\cdot}00051$ cm^{-1}, which is to be compared with Herzberg's (1945) determination, from other infra-red bands, of $q = 0{\cdot}00034$ cm^{-1}.

Data relating to the 13 bands of CO_2 observed in the near infra-red solar spectrum are given in Table 4. The first column gives the designation of each

Table 4. CO_2 Bands in the Infra-Red Solar Spectrum

(The dashes refer to bands predicted but not observed due to obscuration by other bands of CO_2 and by H_2O.)

Desig.	Transition	ν_0 $^{12}C^{16}O_2$	a	b	c
2·06 μ	0 0$_0$ 0 — 0 4$_0$ 1	4853·48	0·77926	−0·00209	$-6{\cdot}2\times10^{-7}$
2·01 μ	0 0$_0$ 0 — 1 2$_0$ 1	4977·78	0·77722	−0·00371	$-7{\cdot}5\times10^{-7}$
1·96 μ	0 0$_0$ 0 — 2 0$_0$ 1	5099·64	0·77905	−0·00267	$-7{\cdot}8\times10^{-7}$
1·65 μ	0 0$_0$ 0 — 0 6$_0$ 1	6075·98	0·78065	−0·00138	$-2{\cdot}2\times10^{-6}$
1·60 μ	0 0$_0$ 0 — 1 4$_0$ 1	6227·90	0·77614	−0·00357	$-1{\cdot}8\times10^{-7}$
1·57 μ	0 0$_0$ 0 — 2 2$_0$ 1	6349·56	0·77640	−0·00371	$-4{\cdot}1\times10^{-7}$
1·54 μ	0 0$_0$ 0 — 3 0$_0$ 1	6503·08	0·77963	−0·00212	$-3{\cdot}6\times10^{-7}$
2·08 μ	0 1$_1$ 0 — $\mathcal{J}$ even — 0 5$_1$ 1	4807·62	0·77920	−0·00184	$-4{\cdot}0\times10^{-7}$
	0 1$_1$ 0 — $\mathcal{J}$ odd	4807·57	0·77737	−0·00262	$-1{\cdot}0\times10^{-6}$
2·01 μ	0 1$_1$ 0 — $\mathcal{J}$ even — 1 3$_1$ 1	4965·37	0·78000	−0·00338	$-6{\cdot}9\times10^{-7}$
	0 1$_1$ 0 — $\mathcal{J}$ odd	4965·35	0·77870	−0·00371	$-3{\cdot}2\times10^{-7}$
—	0 1$_1$ 0 — 2 1$_1$ 1	—	—	—	—
		$^{13}C^{16}O_2$			
2·10 μ	0 0$_0$ 0 — 0 4$_0$ 1	4748·06	0·77980	−0·00142	$-2{\cdot}0\times10^{-7}$
2·05 μ	0 0$_0$ 0 — 1 2$_0$ 1	4887·30	0·77716	−0·00344	$-3{\cdot}5\times10^{-6}$
—	0 0$_0$ 0 — 2 0$_0$ 1	—	—	—	—
		$^{12}C^{16}O^{18}O$			
2·09 μ	0 0$_0$ 0 — 0 4$_0$ 1	4791·22	0·73140	−0·00249	$-3{\cdot}8\times10^{-6}$
2·04 μ	0 0$_0$ 0 — 1 2$_0$ 1	4904·75	0·73128	−0·00324	$-4{\cdot}3\times10^{-6}$
—	0 0$_0$ 0 — 2 0$_0$ 1	—	—	—	—

band, and the second column the quantum numbers involved in the transitions. The third, fourth, fifth and sixth columns give the constants of the least squares solutions of the equation representing the rotational structures, viz.:

$$\nu = \nu_0 + am + bm^2 + cm^3, \qquad \ldots\ldots (1)$$

where $m = -J$ for the negative branch and $m = J + 1$ for the positive branch. For all transitions from the ground state of $^{12}C^{16}O_2$ and of $^{13}C^{16}O_2$, only even values of J are permitted. For the difference bands of $^{12}C^{16}O_2$ and those of $^{12}C^{16}O^{18}O$ both even and odd values of J are allowed.

(ii) *Methane*

Spectroscopic evidence for the presence of methane in the earth's atmosphere was first reported by Migeotte (1948) from observations of the P-branch of the

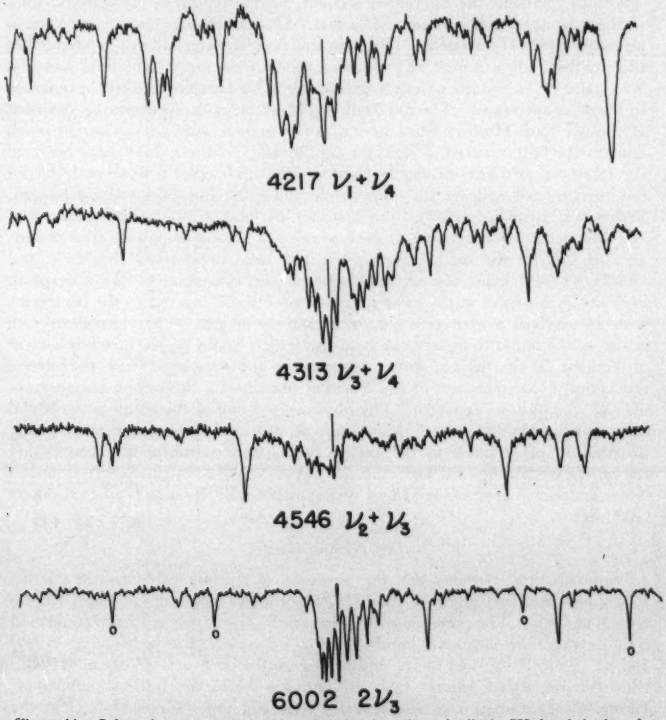


Figure 14. Q-branches of overtone and combination branches of telluric CH_4 bands in the solar spectrum. The band centres are in cm^{-1}.

ν_3 fundamental at $3{\cdot}3\,\mu$ with grating and thermocouple. Independent confirmation was afforded by the detection of the first overtone, $2\nu_3$, at $1{\cdot}66\,\mu$ by McMath, Mohler and Goldberg (1948) with the PbS cell. The latter also succeeded in observing the Q-branches of the combination bands $\nu_1+\nu_4$, $\nu_3+\nu_4$, and $\nu_2+\nu_3$ at $2{\cdot}20\,\mu$, $2{\cdot}32\,\mu$ and $2{\cdot}37\,\mu$, respectively (see Figure 14). Later, Mohler and

Pierce (1949) employed a refrigerated Cashman PbS cell to map the region 2·9–3·6 μ with considerably improved resolution, sufficient to separate several lines of the R-branch of the ν_3 band and to reveal the multiplicity characteristic of the high series members (see Figure 15). Adel (1949) has also reported the presence of the P-branch of ν_3 on records obtained at the Lowell Observatory in 1941.

Recently McMath, Mohler, Pierce and Goldberg (1949) have detected additional faint bands of CH_4 in the solar spectrum at 1·71 μ, 1·73 μ and 2·43 μ. The four bands in the 2 μ region account for approximately 330 telluric lines in the solar spectrum between 2·15 μ and 2·45 μ on tracings obtained at various air masses from 1 to about 8. It is estimated that between one-third and one-half this number of lines would be present on noonday tracings. The weak bands at 1·71 μ and 1·73 μ require a path length of about 3 air masses for their appearance in the solar spectrum. The $2\nu_3$ band of CH_4 in the solar spectrum, as recorded at the McMath–Hulbert Observatory, is shown in Figure 14. The Q-branch is seen to be fully resolved. McMath, Mohler and Goldberg (1949) have analysed the rotational structure of the $2\nu_3$ band and have obtained a value of 5·163 for the average half spacing B_0. The same value has been derived by Nelson, Plyler and Benedict (1948) from analyses of laboratory tracings secured at Northwestern University with a Cashman cell and Wilson amplifier. The abundance of CH_4 in the earth's atmosphere has been determined (McMath and Goldberg 1949) from comparisons between the intensities of the absorption lines of the $2\nu_3$ band in the solar spectrum with those secured in the laboratory with an artificial source and a known quantity of gas. The absorption with zenith sun is found to be equivalent to that which would be produced by a layer of methane 1·1 cm. thick at normal temperature and pressure. From the relative intensities of the members of the P and R branches, a Boltzmann temperature of −33° c. has been deduced. The physical meaning of the temperature found must remain doubtful until knowledge is obtained concerning the vertical distribution of methane in the earth's atmosphere. Information concerning the vertical distribution will be forthcoming as a result of abundance determinations which are now in progress at Mount Wilson, where the elevation is approximately 6,000 feet.

(iii) *Nitrous Oxide*

Spectroscopic evidence for the presence of nitrous oxide in the earth's atmosphere was first given by Adel (1939) through recognition of the ν_1 fundamental at 7·8 μ. The presence of atmospheric N_2O has recently been confirmed by observation of additional bands at 3·9 μ, 4·5 μ and 8·6 μ by Migeotte (1949) and at 3·9 μ, 4·1 μ and 4·5 μ by Shaw, Sutherland and Wormell (1948). Observations with Cashman cells at both the McMath–Hulbert and Mount Wilson Observatories have also revealed four additional bands of N_2O in the solar spectrum between 2·0 and 3·0 μ (McMath and Goldberg 1949). Three of the bands have previously been observed in the laboratory by Plyler and Barker (1931, 1932), namely, $2\nu_1+\nu_3$ at 2·13 μ, $2\nu_3$ at 2·26 μ, and $2\nu_2+\nu_3$ at 3·0 μ. The appearance of the 2·13 μ band in the solar spectrum is shown in Figure 12. The fourth band, which is extremely weak, lies at 2·16 μ and was found on a Mount Wilson tracing obtained with the sun just above the horizon. The identification of the band was established from a high resolution laboratory

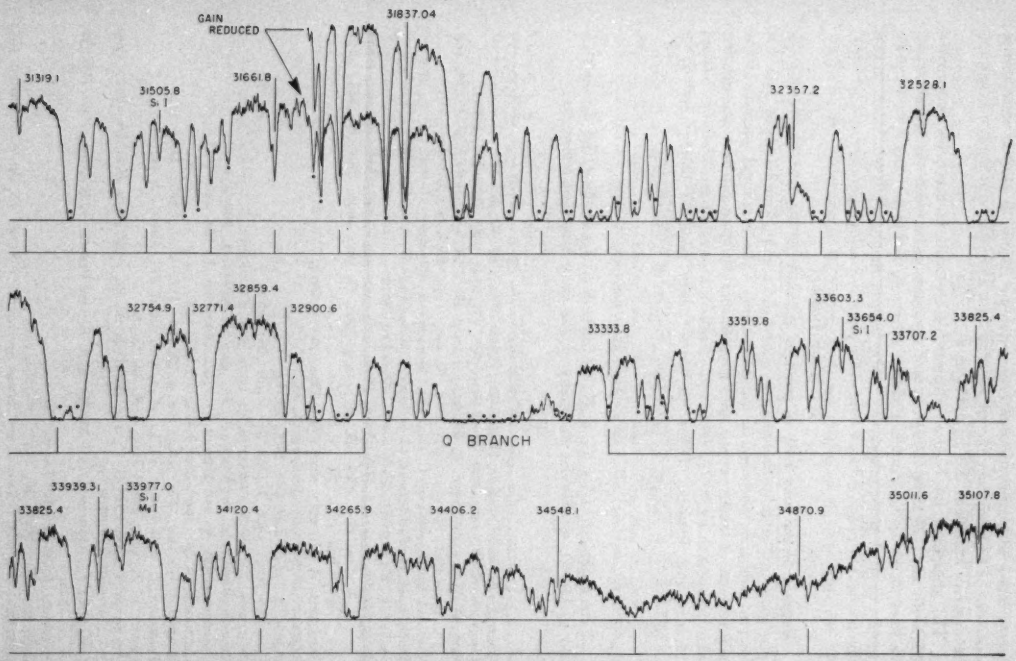


Figure 15. The solar spectrum from 3·1μ to 3·5μ as recorded at the McMath–Hulbert Observatory. The dots indicate lines of H_2O and the vertical lines members of the ν_3 fundamental of CH_4, first identified in the solar spectrum by Migeotte. Solar lines of Si I and Mg I are also shown.

tracing made with an absorption cell containing 25 cm. of N_2O at atmospheric pressure. The laboratory record also discloses a second faint band overlapping the negative branch of the $2\nu_3$ overtone at 2·26 μ.

The abundance of N_2O in the earth's atmosphere has been determined (McMath and Goldberg 1949) from a comparison between the intensities of the lines of the 2·13 μ band in the solar spectrum and as produced by a known amount of N_2O in the laboratory. The telluric absorption is equivalent to that which would be produced by 4 mm. of the gas at normal temperature and pressure.

(iv) *Water Vapour*

Because of the long path of sunlight through the earth's atmosphere, the solar spectrum contains numerous overtone and combination bands of H_2O that are not ordinarily observed in the laboratory. The rotational analysis of the telluric lines of water vapour appearing on the McMath–Hulbert tracings has been undertaken by W. S. Benedict. Preliminary results have been published (Mohler and Benedict 1948) for three newly resolved bands, $2\nu_2+\nu_3$ at 6,871·37 cm^{-1}, $\nu_1+2\nu_2$ at 6,775·00 cm^{-1}, and $3\nu_2$ at 4,666·70 cm^{-1}.

ACKNOWLEDGMENTS

I am indebted to Dr. G. B. B. M. Sutherland for reading the manuscript, and to Dr. O. C. Mohler for many of the early historical references.

REFERENCES

Abbot, C. G., and Freeman, H. B., 1929, *Smith. Misc. Coll.*, **82**, No. 1.
Abney, W. de W., 1880, *Phil. Trans. Roy. Soc.*, **171**, 653; 1886, *Ibid.*, **177**, 457.
Adel, A., 1939, *Astrophys. J.*, **90**, 627; 1941 a, *Ibid.*, **94**, 451; 1941 b, *Ibid.*, **94**, 449; 1942, *Ibid.*, **96**, 239; 1943, *Ibid.*, **97**, 190; 1949, *Phys. Rev.*, **75**, 322.
Adel, A., and Lampland, C. O., 1938, *Astrophys. J.*, **87**, 198.
Babcock, H. D., and Moore, C. E., 1947, *Carnegie Inst. Wash. Publ.*, 579.
Barker, E. F., and Wu, T. Y., 1934, *Phys. Rev.*, **45**, 1.
Becquerel, A. E., 1843, *Ann. Chim. Phys.*, Ser. 3, **9**, 314; 1848, *Ibid.*, **22**, 344; 1869, *C.R. Acad. Sci., Paris*, **69**, 995; 1873, *Ibid.*, **77**, 302; 1876, *Ibid.*, **83**, 249; 1877, *Ann. Chim. Phys.*, Ser. 5, **22**, 360.
Becquerel, A. H., 1883, *Ann. Chim. Phys.*, Ser. 5, **30**, 5.
Brackett, F. S., 1921, *Astrophys. J.*, **53**, 121.
Burns, K., 1920, *Lick Obs. Bull.*, **10**, 64.
Cashman, R. J., 1946, *J. Opt. Soc. Amer.*, **36**, 356 A.
Chalonge, D., and Kourganoff, V., 1946, *Ann. d'Astrophys.*, **9**, 69.
Chandrasekhar, S., and Breen, F. H., 1946, *Astrophys. J.*, **104**, 430.
Chandrasekhar, S., and Münch, G., 1946, *Astrophys. J.*, **104**, 446.
Dennison, D. M., 1940, *Rev. Mod. Phys.*, **12**, 175.
Draper, J. W., 1843, *Phil. Mag.*, Ser. 3, **22**, 360; 1881, *Ibid.*, Ser. 5, **11**, 157.
Foucault, L., and Fizeau, H. L., 1847, *C.R. Acad. Sci., Paris*, **25**, 447.
Fowle, F. E., 1917, *Smith Misc. Coll.*, **68**, No. 8.
Goldberg, L., Mohler, O. C., and McMath, R. R., 1948, *Phys. Rev.*, **74**, 1881; 1949, *Astrophys. J.*, **109**, 28.
Goldberg, L., Mohler, O. C., McMath, R. R., and Pierce, A. K., 1949, *Phys. Rev.*, **76**, 1848.
Herschel, J., 1840, *Phil. Trans.*, **130**, 1.
Herschel, W., 1800, *Phil. Trans.*, **90**, 284, 437.
Herzberg, G., 1945, *Infra-Red and Raman Spectra* (New York : D. Van Nostrand), p. 395; 1948, *Astronom. J.*, **54**, 75.
Higgs, G., 1898, *Astrophys. J.*, **7**, 86.

HUMPHREYS, C. J., and PLYLER, E. K., 1949. (Reported at June Symposium on Molecular Structure, held at Ohio State University.)
KUIPER, G., WILSON, W., and CASHMAN, R. J., 1947, *Astrophys. J.*, **106**, 246.
LAMANSKY, S., 1872, *Phil. Mag.*, Ser. 4, **43**, 282.
LANGLEY, S. P., 1881, *Proc. Amer. Acad. Sci.*, **16**, 342; 1900, *Ann. Smith. Inst.*, 1.
MCMATH, R. R., and GOLDBERG, L., 1949, *Proc. Amer. Phil. Soc.*, **93**, 362.
MCMATH, R. R., and MOHLER, O. C., 1948, *Pub. Astr. Soc. Pac.*, **60**, 119; 1949, *J. Opt. Soc. Amer.*, **39**, 903.
MCMATH, R. R., MOHLER, O. C., and GOLDBERG, L., 1948, *Phys. Rev.*, **73**, 1203; 1949, *Astrophys. J.*, **109**, 17.
MCMATH, R. R., MOHLER, O. C., PIERCE, A. K., and GOLDBERG, L., 1949, *Phys. Rev.*, **76**, 1533.
MEGGERS, W. F., 1918, *Astrophys. J.*, **47**, 1.
MIGEOTTE, M., 1945, *Thèse*, Liége; 1948, *Phys. Rev.*, **73**, 519, *Astrophys. J.*, **107**, 400; 1949, *The Atmospheres of the Earth and Planets* (Chicago : University of Chicago Press), p. 284.
MINNAERT, M., MULDERS, G. F. W., and HOUTGAST, J., 1940, *A Photometric Atlas of the Solar Spectrum* (Amsterdam : Schnabel, Kampert and Helm).
MOHLER, O. C., and BENEDICT, W. S., 1948, *Phys. Rev.*, **74**, 702.
MOHLER, O. C., MCMATH, R. R., and GOLDBERG, L., 1949, *Phys. Rev.*, **75**, 520.
MOHLER, O. C., and PIERCE, A. K., 1949, *Pub. Astr. Soc. Pac.*, **61**, 221.
MOORE, C. E., 1949, *Atomic Energy Levels*, Nat. Bur. Stand. Circ. 467.
NELSON, R. C., PLYLER, E. K., and BENEDICT, W. S., 1948, *J. Res. Nat. Bur. Stand., Wash.*, **41**, 615.
NIELSEN, H. H., 1949. (Reported at June Symposium on Molecular Structure, held at Ohio State University.)
NOBILI, L., 1831, *Pogg. Ann.*, **24**, 640.
PIERCE, A. K., 1949, *Pub. Astr. Soc. Pac.*, **61**, 217.
PLYLER, E. K., and BARKER, E. F., 1931, *Phys. Rev.*, **38**, 1827; 1932, *Ibid.*, **41**, 369.
RUSSELL, H. N., MOORE, C. E., and WEEKS, D. W., 1944, *Trans. Amer. Phil. Soc.*, **34**, Part II.
ST. JOHN, C. E., MOORE, C. E., WARE, L. W., ADAMS, E. F., and BABCOCK, H. D., 1928, *Carnegie Inst. Wash. Publ.*, 396.
SEEBECK, T. J., 1826, *Pogg. Ann.*, **6**, 133, 263.
SHAW, J. H., SUTHERLAND, G. B. B. M., and WORMELL, T. W., 1948, *Phys. Rev.*, **74**, 978.
STARKIEWICZ, J., SOSNOWSKI, L., and SIMPSON, O., 1946, *Nature, Lond.*, **158**, 28.
SUTHERLAND, G. B. B. M., BLACKWELL, D. E., and FELLGETT, P. B., 1946, *Nature, Lond.*, **158**, 873.
SVANBERG, A. F., 1857, *Pogg. Ann.*, **84**, 411.
WILDT, R., 1939, *Astrophys. J.*, **89**, 295.

SHOCK WAVES AND THE PROPAGATION OF FINITE PULSES IN FLUIDS

By W. G. PENNEY and H. H. M. PIKE
Armament Research Establishment, Ministry of Supply

CONTENTS

	PAGE
§ 1. Introduction	46
§ 2. Shock wave relationships : Some numerical values	48
(i) Derivation of the Rankine-Hugoniot equations	48
(ii) The thickness of the shock transition zone	51
(iii) Equations of state and shock wave properties in various fluids	52
§ 3. Reflection, refraction and diffraction of plane shock waves	55
(i) Normal reflection of shock waves by a rigid wall	58
(ii) Oblique reflection at a rigid wall	58
(iii) Reflection at a free surface	60
(iv) Normal reflection at an interface between two fluids	63
(v) Oblique reflection at an interface	64
(vi) Diffraction problems	65
§ 4. Similarity solutions of one independent variable	68
§ 5. Non-similarity solutions	71
(i) A generalization of Riemann's characteristics	71
(ii) Transition of a simple expanding wave into a shock	75
(iii) Calculation of the propagation of shock waves caused by explosions	78
(iv) Physiological effects of blast	81
Acknowledgments	81
References	82

§ 1. INTRODUCTION

RIEMANN'S classical paper on the propagation in one dimension of finite waves in an inviscid, non-thermally conducting fluid is the natural starting point for the presentation of a summary of a large number of theoretical investigations made during the recent war. Riemann proved that if initially a disturbance is created between two planes, $x=a$ and $x=b$, in a fluid elsewhere at rest, the fluid everywhere having constant entropy, then as time increases the disturbance extends in both directions and ultimately consists of two progressive waves, with a central region at rest, one wave moving in the positive x-direction and one moving in the negative x-direction. The solution given is not perfectly general, because certain types of discontinuity in pressure and mass velocity must be excluded, but there is no restriction on discontinuities in the gradient of pressure or mass velocity. The interesting result emerges that progressive waves are not constant in form; speaking descriptively, wherever the pressure gradient, and the related mass-velocity gradient, are negative, facing in the direction of motion of the wave, the gradient becomes steeper, and more negative, as time increases. At some stage the gradient at some point in the waveform becomes infinite and at a later time the pressure–space curve contains a retroverted kink. This kink is physically impossible since it implies three values of the pressure simultaneously over a certain region of space. It is not necessary here to develop Riemann's original solution, nor even to illustrate it by means of a diagram, since a full treatment is given in Lamb's *Hydrodynamics* and in the article by Taylor and Maccoll (1935).

Clearly, as the gradients in fluid velocity and temperature become increasingly steep, viscosity and thermal conductivity can no longer be neglected, and the introduction of these two dissipative mechanisms removes the mathematical dilemma of the retroverted kink, already described. Lamb cautiously observes that Riemann's solution is valid up to the stage where the gradients become infinite at some point, and that at this stage a shock wave is formed.

For the one-dimensional case, therefore, if the pressure is increased above its normal value inside a region of a compressible fluid, a disturbance will be propagated outwards in both directions at finite speeds and at the expanding planes enveloping the disturbance, the pressure, and other physical parameters, will eventually become discontinuous. These surfaces of discontinuity are called shock fronts. There are, however, no corresponding rarefaction shocks (except in certain unusual thermodynamic circumstances), and this point serves to illustrate the non-linear nature of the differential equations governing the propagation of pulses.

No papers have been published giving numerical solutions which show the changing form of a one-dimensional progressive wave and the development of a shock wave, at first infinitesimal, but gradually building up into a finite step. Qualitatively, the situation is clear and the calculations do not appear worth while, but a description of the beginning of the process for a spherically expanding wave, where the situation is less obvious, is given later in this Report.

A number of investigations have been made on finite pulses moving in one dimension, or expanding symmetrically in three, and many of them have a direct connection with problems of blast waves caused by explosions. Throughout the pulse, viscosity and thermal conduction are neglected, and while these two mechanisms are implicitly taken into account at the one place where they are of supreme importance, namely inside the shock front, the necessity of dealing with this complicated boundary region is neatly avoided by the assumption that the shock front is actually a discontinuity of zero thickness. The pressure, mass velocity and entropy on the two sides of the shock front are connected by the Rankine–Hugoniot shock wave equations described later. The continuity of the dynamical and thermodynamical variables inside the pulse, together with the shock wave equations, permit the formulation of the mathematical problems in unique style.

If the equation of state of any fluid and the dependence of its internal energy on the other variables are known, then the thermodynamic properties of any element of the fluid will be completely determined by the values of two independent variables, of which the most convenient pair for use in problems relating to the propagation of pressure pulses are usually the entropy S and either the pressure p or the specific volume v. In a general discussion we must use the most general relationships possible, but three conditions must be satisfied. First, if the fluid can exist under constant pressure conditions, we have

$$(\partial p/\partial v)_S<0, \qquad \ldots\ldots (1)$$

and this is necessary to ensure that the fluid can transmit acoustic waves. Second, it is found experimentally that usually

$$(\partial^2 p/\partial v^2)_S>0, \qquad \ldots\ldots (2)$$

although this may not be true under conditions where negative pressures or tensile stresses can be supported. If (2) is not satisfied, then shock waves of

rarefaction are possible, but not shock waves of compression. Novikov (1948) gives an example where this happens, namely wet steam in the neighbourhood of the critical pressure.

The third condition can be put in many forms, one being

$$(\partial T/\partial v)_S < 0, \qquad \ldots\ldots (3)$$

and this is found to be true with much the same limitations as (2). A fuller discussion is given by Courant and Friedrichs (1948).

In addition to the two thermodynamic parameters describing the fluid element, we need to know the mass velocity, and this will be a vector whose value depends on the system of reference and is, therefore, not an intrinsic property of the element.

§2. SHOCK WAVE RELATIONSHIPS: SOME NUMERICAL VALUES

In this Section we derive the Rankine–Hugoniot equations and explain how numerical values of shock velocity, mass velocity, entropy change, etc., have been calculated as functions of the shock front pressure for various gases and for water.

(i) *Derivation of the Rankine–Hugoniot Equations*

The derivation of the shock wave equations may be found elsewhere, e.g. Taylor and Maccoll (1935) or Lewis and von Elbe (1938). The equations, of course, are familiar and commonly used in work relating to the flow of compressible fluids. Nevertheless, some readers may not be acquainted with the derivation of these equations; it was therefore thought worth while to give them here as well as a few extensions, and to do so in a way that avoided the assumption that the shock front zone was an actual surface of discontinuity.

We disregard the internal mechanics of the shock front except in so far as we assume that in certain circumstances a steady state may be produced. Consider the case of steady one-dimensional flow where the two regions of the fluid, separated by the transition zone, are such that throughout each the conditions are the same everywhere. Inside the zone conditions are also steady, and the frame of reference is chosen to make this zone stay at rest. Use a suffix 0 to denote properties of a fluid element on the low pressure side of the zone, and a suffix 1 to denote properties on the high pressure side. If ρ denotes fluid density, then because fluid is not accumulating anywhere, the rate of mass flow across any section must be constant, say m. Thus

$$m = \rho_1 u_1 = \rho_0 u_0. \qquad \ldots\ldots (4)$$

Figure 1 shows a cylinder of fluid, of unit cross section, containing the shock front zone. Following the motion of this particular mass of fluid, we see that the front, or left-hand edge, of the cylinder is moving with velocity u_1, while the back of the cylinder is moving with velocity u_0. Since the shock compresses the fluid, u_1 must be less than u_0, and the length of the cylinder is therefore decreasing.

The rate of change of momentum of the fluid in the cylinder is simply $m(u_1 - u_0)$, and this must be due to the pressure difference at the two ends, namely $p_0 - p_1$. Hence the equation of momentum is

$$p_0 - p_1 = m(u_1 - u_0). \qquad \ldots\ldots (5)$$

The work done by the cylinder of fluid in unit time is $p_1u_1 - p_0u_0$, and this work appears at the expense of the internal plus kinetic energy of the fluid in the cylinder. Writing E for the internal energy per unit mass, we see that per unit time a mass of fluid m in the cylinder experiences a change of internal energy $m(E_0+\frac{1}{2}u_0^2 - E_1 - \frac{1}{2}u_1^2)$. Hence the equation of conservation of energy is

$$p_1u_1 - p_0u_0 = m(E_0+\tfrac{1}{2}u_0^2 - E_1 - \tfrac{1}{2}u_1^2). \qquad \ldots\ldots (6)$$

The shock wave equations are direct mathematical deductions from (4), (5) and (6). It will be found that

$$E_1 - E_0 = \tfrac{1}{2}(p_0+p_1)(v_0-v_1). \qquad \ldots\ldots (7)$$

This equation is independent of the system of reference.

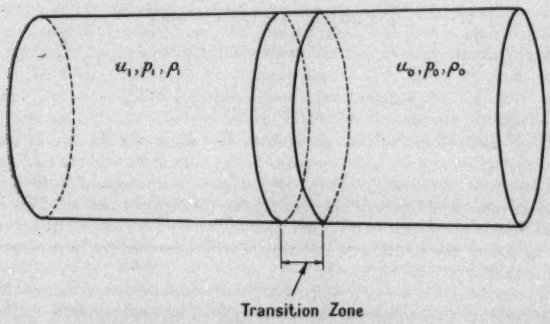


Figure 1. A cylindrical mass of fluid, of unit cross section, passing from right to left through the shock transition zone. Conditions everywhere are assumed stationary, and a system of reference has been taken in which the transition zone is at rest.

The velocity U with which a shock front advances into a stationary fluid is clearly the quantity denoted by u_0 above. Thus

$$U = u_0 = v_0[(p_1-p_0)/(v_0-v_1)]^{\frac{1}{2}}. \qquad \ldots\ldots (8)$$

It can be shown that as $p_1 \to p_0$, $U \to c_0$, where c_0 is the velocity of sound in the fluid at pressure p_0 and density ρ_0.

The mass velocity w behind a shock front advancing into a stationary fluid is the difference between u_0 and u_1, and the direction of w is the same as that of U.

$$w = [(p_1-p_0)(v_0-v_1)]^{\frac{1}{2}}. \qquad \ldots\ldots (9)$$

Equations (7), (8) and (9) are the Rankine–Hugoniot shock wave equations.

Equations (8) and (9) follow solely from the conservation of mass equation (4) and the conservation of momentum equation (5); they therefore apply even to shock waves which cause chemical reactions and release energy, e.g. detonation waves.

For a perfect gas we have $E = pv/(\gamma-1)$, and (7) can be rearranged in the form

$$v_1[(\gamma+1)p_1 + (\gamma-1)p_0] = v_0[(\gamma-1)p_1 + (\gamma+1)p_0].$$

If we write Π for the pressure ratio p_1/p_0, then the compression ratio is

$$\frac{v_0}{v_1} = \frac{(\gamma+1)\Pi+(\gamma-1)}{(\gamma-1)\Pi+(\gamma+1)} < \frac{\gamma+1}{\gamma-1}, \qquad \ldots\ldots (10)$$

so that, no matter how high the pressure in the shock is, the fluid volume is only decreased by a limited factor, less than 6 for $\gamma = 1{\cdot}4$. This limited increase in density implies that the temperature in a strong shock will increase almost in proportion to the pressure instead of in proportion to $\Pi^{(\gamma-1)/\gamma}$, as it would in an adiabatic compression. There is therefore a considerable increase in entropy.

For a perfect gas the entropy is given by

$$S = C_v \ln \mathrm{T} - \boldsymbol{R} \ln p + \text{const.}$$

$$= \boldsymbol{R}(\ln p + \gamma \ln v)/(\gamma-1) + \text{const.}$$

Hence the increase in entropy in the shock compression is

$$S_1 - S_0 = C_v \ln \left[\frac{(\gamma-1)\Pi+(\gamma+1)}{(\gamma+1)\Pi+(\gamma-1)}\,\Pi^{1/\gamma}\right]. \qquad \ldots\ldots (11)$$

For a very weak shock we may write $\Pi = 1+\epsilon$, where ϵ is small. If we then expand the logarithms, we find that terms in ϵ and ϵ^2 cancel, and the leading term is $C_v\{(\gamma^2-1)\epsilon^3\}/12\gamma^3$ or $\boldsymbol{R}\{(\gamma+1)\epsilon^3\}/12\gamma^2$. Hence, to the second order of small quantities, there is no entropy change in going through the shock. This is true for any fluid subject only to the three thermodynamic conditions given earlier. For very strong shocks the entropy change increases roughly in proportion to the logarithm of the shock pressure.

If a pressure pulse, led by a shock front, passes through an element of a fluid, and the pressure then returns to its original value, e.g. atmospheric, the history of the element after the shock front has passed may be taken to be one of constant entropy. Because the shock has increased the entropy of the element, the final temperature of the element is greater than the initial temperature. This temperature excess is quite small in water even for pressures of many tons/in^2, but for strong shocks in air, such as those caused near an explosion, it may amount to thousands of degrees. Energy conservation demands that an amount of energy equal to the increase in enthalpy of the fluid element must have been taken from the pressure pulse. It can be shown that if the pulse retains a constant shape, then all its energy will ultimately be dissipated in this way, with a permanent increase in temperature and volume of every part of the fluid.

The Mach number of the flow in a perfect gas on the high pressure side of the shock front in a system of reference in which the flow on the low pressure side is at rest is

$$M = w/c_1 = \left[\frac{2(\Pi-1)^2}{\gamma\Pi\{(\gamma-1)\Pi-(\gamma+1)\}}\right]^{\frac{1}{2}}, \qquad \ldots\ldots (12)$$

where $c_1 = \sqrt{(\gamma p_1 v_1)}$ is the velocity of sound in the gas at pressure p_1 and specific volume v_1. For a perfect gas the flow behind the shock front is supersonic (i.e. $M>1$) for pressure ratios greater than

$$\Pi_s = 1 + \frac{\gamma(3\gamma-1)+\gamma\sqrt{(13+2\gamma+\gamma^2)}}{4+2\gamma-2\gamma^2}. \qquad \ldots\ldots (13)$$

It will be found that Π_s is 4·60 for $\gamma = 1{\cdot}4$, and 8·85 for $\gamma = 5/3$. As $\gamma \to 2$, $\Pi_s \to \infty$, and for still higher values of γ there are no solutions giving $M > 1$. This suggests that the flow behind a shock in water is always subsonic, since at very high pressures water behaves roughly like a perfect gas with $\gamma = 7$.

(ii) *The Thickness of the Shock Transition Zone*

A number of authors have considered the internal dynamics of the shock transition zone in gases. One group treats the fluid as continuous and assumes that the viscosity and thermal conductivity terms are linear in the velocity and thermal gradients, the two coefficients having their N.T.P. values. Taylor and Maccoll (1935) show that in the case where the pressure ratio on the two sides of the shock is 1·15, the thickness is approximately 3×10^{-4} cm., while for a ratio 8 the thickness is of order 10^{-5} cm. (i.e. just the mean free path in air.). A second group of authors, impressed by the result that the 'continuous fluid' approximation yields shock front thicknesses for intense shocks only equal in length to the mean free path of the gas molecules, attempt solutions more on the lines of classical kinetic theory. Thus, Thomas (1944) improves the classical calculations of Becker (1922), and shows that if the increase of the coefficients of thermal conductivity and viscosity with increasing pressure and temperature are taken into account, the hydrodynamical equations never lead to a thickness small compared with the mean free path. If the shock is strong, the distribution function of molecular velocities within the front departs from Maxwellian. There appear to be no grounds for doubting that classical kinetic theory is capable of proving that all shock fronts in air except the weakest are a few mean free paths thick. The matter, however, does not quite end here, because in the next stage of analysis one must consider the speed of interchange of energy among the various quantized energy states of the individual molecules.

Two independent experimental approaches have indicated that there are 'relaxation times' for the establishment of a Maxwellian distribution among the translational, rotational, vibrational and electronic degrees of freedom, and suggest that vibrations are relatively slow in reaching equilibrium with the others. The first experimental approach is the measurement of the absorption and dispersion of ultrasonic waves. The second approach is the direct observation of spectra of extremely 'hot' radicals in flames. The life-time of the electronic state can sometimes be a few microseconds, and the spectra at low pressure show that the energy distribution among rotations, translations and vibrations is nothing like Maxwellian, while at higher pressure the rotations and translations are in near equilibrium although the vibrations are practically unexcited.

The application of these ideas on relaxation times of vibrational motion to shock waves is not direct, nor is it possible in the present confused state of the experimental evidence to be definite. However, a careful study of the problem has been made by Blackman in a report to the Ministry of Home Security in 1943 (to be published later); he concludes that in a shock wave in pure O_2 at N.T.P. producing a temperature of 5,000° K., the vibrations reach thermal equilibrium at a distance roughly 1 cm. behind the front. For pure N_2 in the same circumstances the distance is between 10 and 100 cm. The presence of minute quantities of water vapour or other non-linear polar molecules will considerably reduce these distances, possibly by a factor as large as 10^5.

Finally, Cowling (1942) has shown that if the fluid passing through a shock front is a mixture of two gases of widely differing molecular weight, then diffusion of heavy molecules toward the high pressure region and light ones toward the low pressure region occurs. This diffusion process may play an important part in the transport of momentum and energy through the shock front and result in an increase in the shock front thickness.

(iii) *Equations of State and Shock Wave Properties in Various Fluids*

The compression ratio produced by a shock is never very large. For water, the ratio is about 1·5 at 50,000 atmospheres, while for air it attains a maximum value of about 12 at 700 atmospheres, decreasing slightly with further increases in shock pressure. For a perfect gas the ratio approaches $(\gamma+1)/(\gamma-1)$ as the pressure increases indefinitely, where γ is the ratio of specific heats on the high pressure side and, therefore, would ultimately approach 4 for any real gas at shock pressures high enough to strip all the electrons from the atoms ($\gamma=5/3$).

At room temperatures the vibrations of the oxygen and nitrogen molecules are practically unexcited, and so for air $C_v=5R/2$, i.e. $\gamma=7/5$. At temperatures of order 3,000° K., the vibrations contribute almost the classical amount to the specific heat and $\gamma=9/7$. At higher temperatures energy is absorbed in excitation of electronic states and molecular dissociation, and γ falls still further. As the number of free atoms increases γ must eventually increase, and in fact γ has a minimum value of 1·18. This value may be used to calculate the shock properties of air with fair accuracy over a range of shock pressures from, say, 200 to 2,000 atmospheres.

If we wish to calculate the shock velocity, mass velocity and the other properties behind the shock front in any fluid as functions, let us say, of the pressure on the high pressure side of the shock, the Rankine–Hugoniot equations give us our answers immediately provided two functions are known. First we require the equation of state; and second, we require the internal energy as a function of two independent thermodynamic variables, say pressure and specific volume. The equation of state of gases may always be taken to be $pv=n\boldsymbol{k}T$, where n is the sum of the numbers of each type of molecule present per unit mass. The justification for this statement is that, as explained above, the compression ratio never exceeds a number of the order of 10. It should be noticed that n is itself a function of T, since strong shocks may raise the temperature sufficiently to cause new types of molecules or dissociated atoms to appear.

The evaluation of the internal energy of a gas over a band of (p, v) values in the vicinity of the shock conditions is straightforward but lengthy. If the shock is weak, only translations and rotations, both classical degrees of freedom, are present, and the perfect gas equations with constant γ apply. With more powerful shocks, vibrational levels and possibly also excited electronic levels appear, and one must evaluate the partition function for these levels, and hence E. Still more powerful shocks cause chemical reactions to occur and the equilibrium concentrations of each class of molecule, each with its own contribution to E measured from some suitable zero, must be calculated. Finally, dissociation and ionization appear, but these two effects can also be handled by the same procedure. Once E has been evaluated in the region of the shock wave conditions for any desired shock front pressure p, a simple iterative process soon gives the required solution of the Rankine–Hugoniot equations.

Penney and Davies (1941), and Brinkley, Kirkwood and Richardson (1944) have independently given the solution of the Rankine–Hugoniot equation for shock fronts in air at pressures from atmospheric up to 1,000 atmospheres (at which shock pressure the temperature is 14,500° K.). These calculations have been extended to shock temperatures of 10^8 deg. K., i.e. shock pressures of 350,000 atmospheres, by Fuchs, Kynch and Peierls (1946). Similar studies have been made by Hirschfelder and Magee (1946).

Sachs (1946) has discussed the properties of very intense shocks where the atoms are completely dissociated into electrons and nuclei and the temperature is so high that most of the pressure is radiation pressure. The shock front is then quite thick, of the order of a mean free path for Compton scattering of photons by electrons. Taub (1948) has obtained the Rankine–Hugoniot equations in relativistic form and has shown that the shock velocity is always less than the velocity of light.

It is interesting to note for shock fronts in air at N.T.P. that the maximum concentration of the molecule NO is caused by a shock front pressure of 180 atmospheres (temperature 5,000° K.). The concentration by mass of this molecule is 12%. Another interesting result which might be mentioned here, and which follows from the papers just quoted and from a knowledge of the shock pressure–distance curve for the explosion of a bare charge of explosive, is that when the air around the explosion comes back to atmospheric pressure, oxides of nitrogen are present to the extent of 0·25–0·50% of the mass of explosive used. These oxides of nitrogen are present because the irreversible heating of the shock leaves some of the air at very high temperature. Turbulent mixing and radiation soon lower the temperature, and the oxides of nitrogen are 'frozen in'.

The shock wave equations for gases other than air may be solved by a procedure similar to that already described. No new principles are involved, but special mention might be made of two cases, each representing an extreme. The first is argon, or equally well any other rare gas, and the second is butane, a gas with a large number of low vibrational frequencies.

Argon is monatomic and, therefore, has no rotational degrees of freedom. Even up to very high temperatures, γ remains at the ordinary value 5/3. The closed-shell electronic configuration of the atom is not easily excited, and the lowest excited electronic state, namely $(3p)^5 4s^3\,P$, lies at 11·7 electron volts above the ground state. At a temperature one-tenth of this amount, viz. 1·7 electron volts, or 13,500° K., the contribution of this electronic level to C_v and C_p is only $0{\cdot}04\boldsymbol{R}$ per mol. It will be found that if the shock raises the temperature from 273° K. to 13,500° K., then the shock pressure calculated assuming $\gamma = 5/3$ (namely 194 atmospheres) is 2% too low, the specific volume is 2% too high, the shock velocity is 0·7% too low and the mass velocity is 1·3% too low. These numerical values indicate the order of the errors, but they are not exact, since other excited electronic levels and ionized states introduce similar, but smaller, corrections.

Butane, CH_3–CH_2–CH_2–CH_3, or its isomer, is a molecule containing 14 atoms. The number of vibrational modes of freedom is therefore 36. Some of these, particularly the torsional modes, have extremely low frequencies (of the order 200 cm^{-1}), while others, the C–H vibrations, have considerably higher values (of the order of 3,000 cm^{-1}). At N.T.P. many of the vibrations should be excited, and this expectation is confirmed by the very low experimental value 1·1 for γ. As the temperature is raised further vibrations are excited, and the

value of γ must fall. If we could treat the vibrational modes as classical, then the γ value would ultimately become 77/75, but before the temperature could be raised sufficiently for this approximation to be valid, the molecule would break up. Experimentally, it is found that butane begins to decompose slowly at 600° c. and sufficiently fast to affect shock waves of normal duration at about 1,000° c.* Without making accurate calculations we can see that shock fronts at any pressure level are much less effective in raising the temperature in butane than they are in air and, *a fortiori*, in argon.

The solution of the shock wave equations for water has been given independently by United States and British authors. Because the work was directed towards naval problems, salt water and pure water were considered. The modification due to the salt and other dissolved material was trivial. Calculations up to 80,000 atmospheres have been published by Richardson, Arons and Halverson (1947). At high pressures they used Tait's equation of state in a slightly modified form,

$$\{p+B(S)\}\{v(p,S)\}^n = B(S)\{v(0,S)\}^n, \qquad \ldots\ldots (14)$$

where n and $B(S)$ were estimated by fitting the observed compressibility data. Actually it was found sufficient to fix n at the single value 7·15, but to make B vary slightly with temperature, or entropy. The function $v(0,S)$ was found by fitting a quadratic expression in T to the thermal expansion data. Almost the identical equation was given by Penney and Dasgupta (1942); the value chosen for n was 7·25. These authors selected the parameters by fitting Bridgman's compressibility data for pure water. The results for the shock velocity, mass velocity and temperature rise obtained by the two groups of authors agreed over the whole range within one or two per cent. It might be noticed that for pressures below 1,500 atmospheres Ekman's (1908) empirical equation of state fits the experimental data very accurately indeed, and in this region is to be preferred to (14).

Table 1 gives a few selected values of the shock wave variables as functions of the pressure on the high pressure side of the shock front. The values for air, argon and water are correct to within about 1%, while those for butane are less accurate as they are based on the perfect gas equations with $\gamma = 1{\cdot}1$. In every case the pressure on the low pressure side has been taken as atmospheric and the temperature as 300° K. The velocities of sound in the various fluids on the low pressure side are 347, 323, 221 and 1,532 m/sec. respectively.

A striking demonstration of the widely different temperatures reached when a given strong shock passes through argon, air and butane has been made by Marley (1942). The results are shown in Plate I. A high speed metallic jet was made from a small-shaped charge coated with a thin metallic liner. The jet was directed through a box containing argon, air and butane in various compartments separated by thin Cellophane walls. A photograph was taken and the contrast in the relative luminosities in the bow waves of the jet in the three gases is very apparent. The extra luminosity as the bow wave in argon and air strikes the Cellophane membrane is also clearly seen.

Finally, mention might be made here of shocks in foam-like and powder materials. Very large variations in the individual circumstances are possible, but two limiting cases of some practical importance exist. The first is that in

* These temperatures were given to the authors by Dr. A. J. B. Robertson.

which the heavy component of the mixture has little intrinsic strength and may be considered incompressible, while the light component is a gas having negligible mass but considerable compressibility (examples are underwater blast waves impinging on cork or sorbo-rubber, air blast and underwater blast waves passing into the lungs of men or animals: see the concluding paragraphs of §5). The second case is where the shocks are so enormous that the heavy component itself is compressed and may be treated as a liquid in the compressed state (an example is an intense shock wave sent into a metal powder from the detonation of a high explosive). It seems that no attempt has been made to investigate the

Table 1. Shock Wave Properties in Various Fluids

Fluid	Pressure (atmos.)	Shock velocity (m/sec.)	Fluid velocity (m/sec.)	Sonic velocity (m/sec.)	Temperature rise in shock (° c.)	Temperature rise when reduced adiabatically to one atmosphere (° c.)
Air	20	1430	1140	680	900	250
	100	3130	2730	1110	3440	1240
	200	4370	3920	1470	5090	2350
	500	6850	6270	2160	7940	3860
Argon	20	1300	910	790	1480	240
	100	2890	2140	1640	7480	930
	200	4080	3040	2300	14980	1530
Butane	20	970	875	310	285	145
	50	1530	1430	405	710	410
Water	2000	1760	110	1880	5	<0·02
	10000	2335	426	2755	36	5
	20000	2880	690	3535	70	18
	50000	4020	1235	5100	152	52

One atmosphere has been taken as one bar or 14·50 lb/in².

shock-transition zone in materials such as those just mentioned; nor is it clear, at least in the first case, that a 'steady state' transition, such as that assumed earlier in deriving the Rankine–Hugoniot equations, always exists, or ever exists.

It would appear that a shock wave consisting of a very thin front leading a pulse of constant pressure is not an adequate representation. The thickness of the front must be at least the 'shock velocity' multiplied by the time taken for the air in the cavities to cool into thermal equilibrium with the walls.

§3. REFLECTION, REFRACTION AND DIFFRACTION OF PLANE SHOCK WAVES

In the following paragraphs we consider the principal results which are known about reflection, refraction and diffraction of plane shock waves. There is disappointingly little to report about diffraction, mainly because of the complexity of the mathematical equations.

Certain cases relating to the reflection and refraction of shock waves are particularly simple because the shock wave pattern can be reduced to a steady state merely by superposing a mass velocity on the whole system. For these cases the experimental evidence supports the mathematical results. Other

cases, at first sight equally simple, are such that the mathematical equations show that steady state solutions are not possible, although similarity solutions (as illustrated by an example below) appear to exist. A study of these cases in an idealized form throws light on a wide range of experimental observations, but some points of detail have not yet been fully elucidated, possibly because the experiments do not exactly conform to the situation envisaged in the mathematical treatment.

Imagine a plane shock wave passing through a fluid, and suppose that the pressure, mass velocity and entropy are constant everywhere behind the shock. Let the shock be approaching an infinite wedge, whose leading edge (Figure 2 (*a*)) is parallel to the plane of the shock.

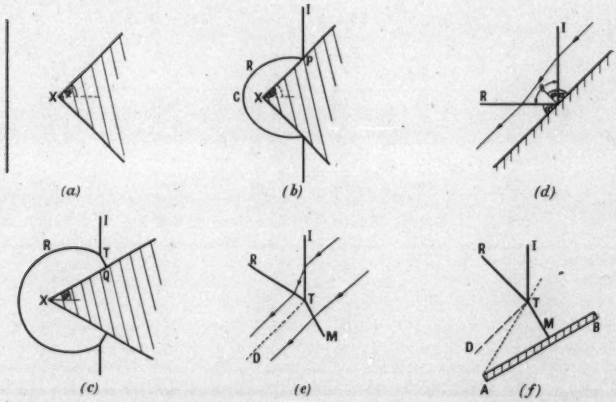


Figure 2. (*a*) A plane shock front, behind which the pressure is constant, approaching a wedge. (*b*) One type of diffraction pattern. (*c*) The second type of diffraction pattern. (*d*) A stationary reflection pattern (i.e. shock fronts at rest by superposing a mass velocity on the whole system), representing the idealization of conditions in (*b*) near P. (*e*) The intersection pattern of three shock fronts meeting at a point, representing the idealization of conditions in (*c*) near T. (*f*) The growth of the Mach Y when a shock front strikes a finite wall, showing the slipstream D.

When the shock has passed X, a diffracted shock wave has been formed. The diffracted pattern will be similar in shape at all times, and the linear dimensions of the pattern will be directly proportional to the time that has elapsed since the initial shock reached X. In other words, we are considering a problem exhibiting *similarity*. The time variable t can be eliminated from the mathematical equations by taking $(x/t, y/t)$ as new independent variables, where x and y are the usual Cartesian coordinates, or in some cases the Lagrangian coordinates, specifying the initial position of a particle. The principle of similarity thus reduces the number of independent variables from three (viz. x, y and t) to two. Because the time variable disappears in these new variables, the problem is sometimes called pseudo-stationary.

If, without having studied these matters, one were asked to sketch the trace of the diffraction pattern, one would probably produce a sketch like Figure 2 (*b*).

Photograph of luminosity along tracks of fragments moving at 15,000 ft/second through cellophane compartments containing air, butane, argon and air respectively (Marley).

(This photograph is reproduced by permission of the Controller, H.M. Stationery Office.)

PLATE I.

To face page 56

This photograph is a print of one-half of a plate showing the shock from a shallow underwater explosion. A spark was triggered to give the shape of the shock at any desired time. The shock front away from the water surface is accurately spherical, but near the surface the front bends back several degrees. The exact amount of the bending back depends on the depth of the charge and the shock front pressure. This effect is in some ways analogous to the formation of the Mach Y shown in Figure 2 (McEnhill).

(*This photograph is reproduced by permission of the Controller, H.M. Stationery Office.*)

PLATE II.

At the point P the incident wave gives rise to a reflected wave, and the reflected wave curves round somewhat as shown. The length XC would be less than XP, because C is moving against the stream, while P is moving with it.

When the incident shock is very weak (i.e. an acoustic wave), the angle of reflection at P would equal the angle of incidence, while the curved surface would be part of a circle with centre X. The complete solution of the diffraction problem in this case, as well as for cases of more complicated types of acoustic pulses, has been given by Friedlander (1946).

Figure 2 (*b*) represents the experimental facts provided the semi-angle of the wedge is between 90° and an angle ϕ_1, where ϕ_1 is approximately 50° for strong shocks and tends to zero for weak shocks. When the semi-angle of the wedge is less than ϕ_1, shadow and spark photographs show that the shock wave system is like that shown in Figure 2 (*c*). Instead of the trace of the incident and reflected waves meeting at a point on the surface of the wedge, as at P in Figure 2 (*b*), a single slightly curved shock leaves the surface of the wedge normally at Q. This shock runs into a triple line T, one branch being the initial shock and the other forming a curved boundary around the apex of the wedge. The type of reflection observed at Q is one example of 'Mach reflection'. The trace of three shocks meeting at a point is often described as a 'Mach Y', while the single shock surface leaving the surface at Q is called the 'Mach stem'.

Many photographs have been taken of diffraction patterns of types similar to those described above. The complete diffraction problems are too difficult to solve completely; the difficulty arises from the fact that a shock front travels at subsonic speed relative to the compressed fluid behind it, so that disturbances originating behind the front will overtake it. Conditions at any point inside the curved reflected shock front therefore affect and are affected by other points inside that front. However, if the curvature of the reflected shock front in Figure 2 (*b*) does not increase greatly as we approach P, then inside any sufficiently small region containing P we may regard the reflected shock front as plane. The reflection pattern at P in Figure 2 (*a*) can then be understood in terms of the idealization shown in Figure 2 (*d*). Here an infinite plane shock, with constant 'backing', is incident on a rigid wall. An infinite reflected plane shock is also present, and the shock pattern can be reduced to a steady state. A typical stream line, similar in geometrical form to all others, is drawn. The case represented in Figure 2 (*d*) is called 'regular oblique reflection'. The Mach Y in Figure 2 (*c*) leads one to study the three infinite plane shock problem sketched in Figure 2 (*e*). In this case the system of reference may be chosen such that one shock is reduced to rest and that the stream lines through this shock are perpendicular to the shock front.

There are two obvious generalizations of the case shown in Figure 2 (*d*). In the first, the rigid plane wall, and the space below, are filled with a second medium. An infinite plane refracted shock then extends through the second medium, and in certain cases, depending on the 'acoustical impedance' of the second medium, a reflected shock will be formed, while in others there will be no shock but a rarefaction zone. In the second generalization the rigid wall of Figure 2 (*d*) is replaced by a 'free surface'. The medium in this case can only be a liquid.

The following paragraphs develop these various cases in slightly greater detail.

(i) *Normal Reflection of Shock Waves by a Rigid Wall*

If we wish, we may regard the rigid wall as replaced by a mirror image of the incident shock and study the intersection of the two, but this idea introduces no essential simplification. The wall imposes the condition that the fluid shall have zero velocity normal to its surface, and it achieves this result by sending a reflected shock back through the oncoming fluid. The reflected shock must be of such strength that it reduces the oncoming fluid to rest. For weak shocks this implies that the excess pressure in the reflected shock equals that in the incident shock, i.e. that reflection doubles the excess of pressure above that formerly existing in the undisturbed fluid.

For a perfect gas the ratio Π' of the absolute pressure in the reflected shock to that in the incident shock is given in terms of Π, the ratio of the absolute pressure in the incident shock to the undisturbed pressure by

$$\Pi' = \frac{(3\gamma-1)\Pi-(\gamma-1)}{(\gamma-1)\Pi+(\gamma+1)}. \qquad \ldots\ldots (15)$$

We see that this gives the expected pressure doubling for very weak shocks but that for strong shocks the pressure is much more than doubled.

For the reflection of strong shocks in air Penney and Davies (1941) have shown that Π' has a maximum value of 14 for $\Pi = 500$. This agrees quite well with the value calculated from our 'high pressure' value of γ for air, namely 1·18. For water, the equation of state suggests that the natural unit of pressure is B, sometimes called the 'breakdown pressure' and equal to about 3,000 atmospheres. Penney and Dasgupta (1942) have shown that for pressures less than B the increase in pressure on reflection is almost the same as for acoustic waves, Π' being 2·17 for an incident shock pressure of 1,000 atmospheres, rising to 2·9 at 10,000 atmospheres and 3·9 at 50,000 atmospheres incident shock pressure. At such pressures no real wall remains rigid, but the equivalent physical process of the interaction of two strong shocks can still take place.

(ii) *Oblique Reflection at a Rigid Wall*

A theoretical study of the reflection of plane shock waves has been made by von Neumann (1943), and detailed numerical results for regular reflection for both perfect gases and water-like substances are given by Polachek and Seeger (1944). Smith (1945) has used a shock tube to study Mach reflection. Parallel theoretical work was carried out in Germany by Weise (1943). Courant and Friedrichs (1948) devote a whole chapter to the subject.

When the plane of the incident wave makes an angle α, and the reflected wave an angle α', with the reflecting plane, it is the components, normal to the reflecting surface, of the fluid velocities produced by the two shocks that must be equal and opposite. The angle α is the complement of the semi-angle ϕ of the wedge mentioned earlier. The reflection is called regular when the incident and reflected shock fronts intersect in the surface, and this imposes a second condition. These two conditions are sufficient to determine both the strength and direction of motion of the reflected shock provided it is plane, so that we can solve the reflection problem completely for a small region near the point P shown in Figure 2(*b*).

Between the two shock fronts there is fluid that is compressed to the incident shock pressure. Relative to that fluid, the incident shock is travelling with subsonic

speed σ and the reflected shock with supersonic speed σ'. For convenience, we take our unit of velocity to be sonic speed in the once compressed fluid, so that $\sigma' > 1 > \sigma$. Polachek and Seeger show that for a perfect gas σ'^2 must satisfy a quadratic equation whose coefficients are functions of σ, γ and α. This equation has two solutions giving either a reflected shock of strength almost equal to that for the head-on case and making an angle α' with the plane a little smaller than α but measured in the opposite sense, or else a very strong shock with its front nearly normal to the reflecting plane. The first of these two solutions is the one actually observed. As α is increased, the two solutions for the reflected wave approach each other and ultimately coincide for a value of α near 40° for incident shock pressures in air greater than two atmospheres absolute. For weak shocks the critical angle increases toward 90°. For water much the same occurs except that for extremely strong shocks the limiting value of α is 15° and that for incident shock pressures of a few thousand atmospheres, which are weak on our natural scale of pressures, the limiting value of α varies considerably, being 51° for 1,000 and 29° for 10,000 atmospheres.

One other point of interest is that for an incident shock of given (high) pressure in air the reflected shock pressure diminishes with increasing α to a minimum of just over 80 per cent of the head-on value for α near 35°, and then rises steeply as α approaches the critical value. On the other hand, shocks in water behave like weak shocks in air, the reflected pressure increasing continuously with α. For incident shock pressures of a few thousand atmospheres this increase amounts to some 60 per cent when α reaches its critical value.

A physical explanation of the breakdown of regular reflection for sufficiently large values of α has been given by Cole (1948) on the following lines.

Let the incident and reflected shocks have velocities σ and σ' relative to the intermediate fluid, which is travelling with velocity w along the normal to the incident shock. Writing θ for $(\pi - \alpha - \alpha')$, the angle between the two shocks, then the reflected shock front has a velocity $\sigma' + w\cos\theta$ along its own normal. The velocity along the reflecting surface of the point of intersection of the incident shock is $(\sigma + w)\operatorname{cosec}\alpha$ and that of the reflected shock $(\sigma' + w\cos\theta)\operatorname{cosec}\alpha'$.

Now as α increases so does α', and $\theta \to 0$; since $\sigma' > \sigma$, a value of α must eventually be reached such that the second velocity is greater than the first, so that O, the point of intersection of I and R, must rise above the reflecting surface. We then have Mach reflection.

Physical intuition would suggest that for a given incident shock strength there would be a value of α for which the velocity of O normal to the reflecting surface would be infinitesimal. We should then have either regular reflection or what is called a stationary Mach reflection, with O travelling parallel to the reflecting surface at a distance from it determined by the initial conditions. For all larger values of α we should expect Mach reflection. This picture agrees with experiment for shocks of moderate strength in air although there is a considerable region of overlap, Mach reflection setting in before the theoretical limit of regular reflection is reached. Reflections of very strong shocks have not been measured and there are still some discrepancies between theory and experiment for very weak shocks.

The case when regular reflection breaks down has been studied by Smith using a small reflecting plane AB in a shock tube (Figure 2(f)). In this case the point of intersection T of the incident and reflected waves follows a straight

line starting from the edge of the reflecting plane. Between T and the plane a third wave appears, called the Mach stem. Since reflection is equivalent to interaction with another wave coming from below the plane, we see that by symmetry the Mach stem must intersect the reflecting plane normally. The fluid that passes through the normal shock front M will be at the same pressure but a different entropy from that which passes through the two shocks I and R. There will therefore be a plane of discontinuity or slipstream D across which there will be changes in fluid density, temperature and velocity.

Theoretical work on the intersection of three plane shock waves has been carried out by von Neumann and others, but agreement with Smith's results is unsatisfactory because the physical picture assumed in the mathematics is too simple. The Mach wave appears to be not quite plane, and a number of other possibilities have been suggested. A limited discussion is given by Courant and Friedrichs (1948).

It is known from experiment that the regular type of reflection does occur under some conditions, and the laws governing such reflections are found to hold near the line of intersection of the waves. We still have an elliptic mathematical problem, but the dependence of conditions in a small region round the line of intersection on conditions elsewhere becomes negligible when that region is small. We should likewise expect agreement between theory and experiment for reflection at an interface between two fluids or at a free surface so long as all waves were nearly plane and all intersected in that surface.

In the Mach case calculations are based on the assumption that three plane waves meet in a line T. We are led immediately to postulate a plane of discontinuity D, passing through T, separating fluid which has passed through one shock from that which has passed through two, and which has, therefore, gained a smaller increase in entropy. Fluid pressures will be the same on the two sides of D but densities, and hence velocities, will differ. This implies that on one side or the other the fluid has a component of velocity normal to the wall, and this is an impossible condition. Our picture therefore is too simple; it may be correct for a very small region containing the line T, but outside that region either one of the shock fronts or the plane D must be curved, or else some additional wave system must be present. The difference in velocity across D is due to the difference in entropy. It is therefore somewhat surprising to find that the assumption of an intersection of three plane shock waves gives results agreeing roughly with experiment for shocks of moderate strength but not at all for weak ones.

(iii) *Reflection at a Free Surface*

The boundary condition is that the pressure at the free surface must always be zero. We consider first the case analogous to regular reflection, the incident and reflected waves intersecting in a line in the surface. In Figure 3 this is the line through O normal to the plane of the paper. OA is the surface in front of the incident shock OS and OB the surface behind it. Relative to a stationary coordinate system, the fluid is at rest in front of the shock which is moving into it with velocity U; and behind it the fluid has velocity w in the same direction.

Relative to a coordinate system moving with O the flow must be steady. A fluid element moves along PS with velocity $U \sec \phi_1$, where ϕ_1 is the angle of incidence of the shock. After passing the shock front it is deflected along SR

and now has the velocity u_2 obtained by adding vectorially the velocity w to its previous velocity, w being directed perpendicular to OS. At R it reaches a rarefaction wave coming from O, the angle SRO being the Mach angle α_2 defined by

$$u_2/c_2 = \operatorname{cosec} \alpha_2, \qquad \ldots\ldots (16)$$

where c_2 is the velocity of sound in the compressed fluid behind the shock front. The rarefaction zone will have a finite width and the fluid particles will turn along a curved path as they pass through the zone until the pressure has fallen

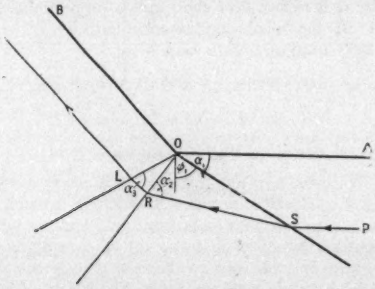


Figure 3. The reflection of a shock front at the free surface of a liquid. The front OS has been reduced to rest by superposing a velocity on the system. A Prandtl–Meyer expansion about O results, provided the angle ϕ_1 is greater than a critical value, depending on the properties of the liquid and the intensity of the shock.

to zero at some point L. The increase in kinetic energy of the fluid element as it passes down the pressure gradient along RL is given by Bernoulli's equation

$$\tfrac{1}{2}(u_3^2 - u_2^2) = -\int_R^L dp/\rho. \qquad \ldots\ldots (17)$$

The direction of flow is now parallel to the surface OB; the angle between the direction of flow at L and the radius vector OL, which equals the angle LOB, is the Mach angle α_3, given by

$$u_3/c_0 = \operatorname{cosec} \alpha_3 \qquad \ldots\ldots (18)$$

where c_0 is the velocity of sound in the fluid at zero pressure. The angle ROL and properties of the fluid along other radii inside that angle can be determined by numerical integration involving only a single quadrature.

A point worth noting is that the diagram is devoid of any given length but A, S, R, L and B are specified by the directions of the radii vectores to them from O. The properties of any element of fluid can therefore only depend on the direction of the radius vector to it from O. Any flow of a type where the properties of the fluid are functions of an angle only is a Prandtl–Meyer flow, and our case is a particular example of this type of flow. A necessary condition for the stability of such a flow, deducible from the hydrodynamic equations, is that the component of velocity of any fluid element perpendicular to its radius vector from O must be sonic.

Another necessary condition for a flow of this kind to exist is that α_2 must be real, i.e. that u_2 be supersonic. For an incident shock of given strength this will always be the case for values of ϕ_1 sufficiently near 90°; but $U-w$ is subsonic relative to the fluid behind the shock and there must always be some value of $\phi_1>0$ for which u_2 is just sonic; for lower values of ϕ_1 our model breaks down. We see that, following a particle at a given depth below the surface passing through the system, the full shock pressure only lasts between the time of arrival at the shock front OS and the time of arrival at the rarefaction wave OR, and thereafter the pressure falls rapidly. For very small depths the duration of the pressure pulse is therefore very short and is proportional to the depth, but during that short time the full shock pressure obtains.

We have for the critical minimum value of ϕ_1

$$c_2^2 = u_2^2 = (U\sec\phi_{cr} - w\cos\phi_{cr})^2 + (w\sin\phi_{cr})^2,$$

$$\text{i.e.} \quad \sec^2\phi_{cr} = (c_2^2 + 2wU - w^2)/U^2. \qquad \ldots\ldots (19)$$

Using an equation of state of the Tait type for water,

$$v = 1{\cdot}163/(2{\cdot}94 + p)^{0{\cdot}138}, \qquad \ldots\ldots (20)$$

where the unit of pressure is 1,000 atmospheres (14,500 lb/in²), to calculate c_2, w and U as functions of the shock pressure, we obtain Table 2.

Table 2. Critical Values of Angle of Incidence for Various Shock Pressures

p (lb/in²)	100	400	1000	4000	10000	30000	70000
ϕ_{cr} (deg.)	2·1	4·2	6·4	15·3	18·0	27·0	38

For $p \ll 40000$ lb/in², $\phi_{cr} \sim 0{\cdot}21\sqrt{p}$.

Underwater explosions generate a spherical shock front, but our model will still serve for determining conditions near the surface. The shock first strikes the surface above the explosion at normal incidence, but as the wave spreads out the angle of incidence diminishes. The pressure in the shock front diminishes and so ϕ_{cr} also diminishes, but there is always some finite radius from the explosion at which ϕ_1 catches up ϕ_{cr}, and our model breaks down. This follows since at large distances both ϕ_1 and p diminish inversely as the radius, but ϕ_{cr} only diminishes as $\sqrt{p}$.

Inside the critical radius our model serves to give a good picture of what happens at points whose depth below the surface is small compared with the length of the pulse. Figure 4 shows a typical pressure–time curve for the shock wave produced by an underwater explosion in deep water, the full line giving the pressure measured at a point near the surface and the broken line the pressure measured at the same radius from the explosion but at a much greater depth below the surface. For completeness it should be mentioned that at great depths the positive pressure pulse from a 40-lb. high explosive charge lasts about a millisecond and is followed by weaker pressure pulses due to oscillations of the gas bubble.

The interesting question of the flow pattern near the surface when the Prandtl–Meyer flow fails will not be treated here although we have made progress on the problem. We have considered a plane shock travelling tangentially along

and under an upward sloping shelf until it strikes obliquely the free surface of the water. A similarity solution may then be constructed with all boundaries limiting the diffraction pattern expanding linearly with time about the line where the shelf meets the surface. The shock front near the surface bends back. There is a narrow region of angles of incidence within which two regimes are possible, corresponding closely to the two types of reflection at a rigid wall.

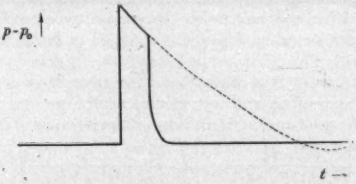


Figure 4. The form of the pressure–time curve for an underwater explosion, showing how a nearby free surface 'cuts off' the pulse.

The bending back of the shock front at the free surface in water when the front is at nearly grazing incidence is shown in Plate II. A very small charge was detonated under water in a tank with glass windows, and a spark was triggered to give a shadow photograph of the shock front at any desired position. The actual amount of bending back of the front near the surface over and above the spherical curvature is small, and in the example shown is about 3°. We are indebted to Mr. J. J. McEnhill for providing us with a plate from which the photograph was copied.

(iv) *Normal Reflection at an Interface between two Fluids*

We saw earlier that a plane shock wave in any fluid of known properties can be specified in terms of one parameter. At an interface between two fluids two conditions must be satisfied, namely continuity of pressure and continuity of fluid velocity across the interface. In addition to the given incident shock, there must therefore be two other waves, a reflected one and a transmitted one, to give the two degrees of freedom necessary to satisfy the boundary conditions.

The case of two perfect gases has been treated in a simple manner by Paterson (1948) on the following lines. Assume in the first place that all three waves are shock waves, the pressure ratios across the incident, transmitted and reflected shocks being Π_1, Π_2 and Π_3 respectively. Then the fluid velocities behind the three shocks, all measured in the direction of propagation of the incident shock, are given by

$$\frac{w_1}{c_1} = \sqrt{\left[\frac{2}{\gamma_1}\right]} \frac{\Pi_1 - 1}{\sqrt{[(\gamma_1+1)\Pi_1 + (\gamma_1-1)]}}, \qquad \ldots\ldots (21)$$

$$\frac{w_2}{c_2} = \sqrt{\left[\frac{2}{\gamma_2}\right]} \frac{\Pi_2 - 1}{\sqrt{[(\gamma_2+1)\Pi_2 + (\gamma_2-1)]}} = \sqrt{\left[\frac{2}{\gamma_2}\right]} \frac{\Pi_3\Pi_1 - 1}{\sqrt{[(\gamma_2+1)\Pi_3\Pi_1 + (\gamma_2-1)]}} \qquad \ldots\ldots (22)$$

and

$$\frac{(w_1 - w_3)}{c_3} = \sqrt{\left[\frac{2}{\gamma_1}\right]} \frac{\Pi_3 - 1}{\sqrt{[(\gamma_1+1)\Pi_3 + (\gamma_1-1)]}}, \qquad \ldots\ldots (23)$$

where c_1 and c_2 are the velocities of sound in the two gases before being disturbed and c_3 is the velocity of sound in that part of the first gas which is compressed by the reflected shock. The condition of the equality of pressures across the

boundary, i.e. behind the transmitted and reflected shocks, has already been used. We now have to make Π_3 fit the other boundary condition, $w_3 = w_2$. Since the incident wave is a shock, $\Pi_1 > 1$, and from the equation (22) we see that w_2 is a monotonic increasing function of Π_3, passing through zero for $\Pi_3 = 1/\Pi_1$. Again, from equation (23), w_3 is a monotonic decreasing function of Π_3, being equal to w_1 and greater than zero for $\Pi_3 = 1$. Hence the solution of the two equations must give $\Pi_3 > 1/\Pi_1$, i.e. $\Pi_2 > 1$, so that the transmitted wave is always a shock. If we consider either the transmitted or reflected wave to be a rarefaction, the corresponding formula is changed in form, but the argument is exactly the same.

If the reflected wave is to be a shock, we must have $\Pi_3 > 1$ or, using the monotonic behaviour of w_2 and w_3, we must have $w_3 > w_2$ for $\Pi_3 = 1$. Since the initial pressures are equal on both sides of the boundary, this gives us

$$\frac{v_1}{(\gamma_1+1)\Pi_1+(\gamma_1-1)} > \frac{v_2}{(\gamma_2+1)\Pi_1+(\gamma_2-1)},$$

where v_1 and v_2 are the initial specific volumes of the two gases. If the two gases have the same value of γ an increase in density, and hence a decrease in sonic velocity, gives rise to a reflected shock, and conversely a decrease must give rise to a reflected rarefaction. For $\gamma_1 \neq \gamma_2$ Paterson puts the condition in the form

$$(\Pi_1-1)\{(\gamma_1+1)\rho_1-(\gamma_2+1)\rho_2\} < 2(\gamma_2\rho_2-\gamma_1\rho_1). \quad \ldots\ldots (24)$$

Since γ and v are independent parameters, four cases can arise: (i) if $\gamma_1\rho_1 < \gamma_2\rho_2$ and $(\gamma_1+1)\rho_1 < (\gamma_2+1)\rho_2$ then the reflected wave is always a shock; (ii) if $\gamma_1\rho_1 > \gamma_2\rho_2$ and $(\gamma_1+1)\rho_1 > (\gamma_2+1)\rho_2$ then the reflected wave is always a rarefaction; in cases (iii) and (iv) the reflected wave is a shock for some incident shock strengths and a rarefaction for others. If we replace (Π_1-1) by the actual pressure difference (p_1-p_0), the right-hand side of (24) must also be multiplied by p_0 and the quantities in the bracket in the right-hand side are the squares of the acoustic impedance ρc of the two gases. Paterson shows that conditions (iii) and (iv) can be satisfied by a few pairs of gases, such as neon and acetylene.

From the reciprocal nature of (i) and (ii) we see that if the direction of propagation of the incident shock across the boundary is reversed then the nature of the reflected wave is changed.

(v) *Oblique Reflection at an Interface*

The algebraic treatment of a shock wave impinging obliquely on the interface of two media is complicated, although the principles are straightforward. The cases which give rise to a regularly reflected shock wave have been studied by Taub (1947). The angles of reflection and refraction must be treated as variables, as well as the strength of the reflected and transmitted shocks. Geometrical and dynamical relationships are just sufficient to fix all unknowns in terms of the angle of incidence and the strength of the incident shock, but the elimination of all but one unknown, thereby obtaining an equation for calculating this unknown, is lengthy. Conditions are very much the same as for the three-wave intersection discussed under Mach reflection and, for the reason given there, namely the different entropy change in going to a given pressure through one shock or through two shocks, the boundary will always be a slipstream. In this case the condition (24) is replaced by a relation giving the tangent of the angle τ between the direction

of motion of the incident shock and the boundary in terms of the four other parameters Π_1, γ_1, γ_2 and ρ_1/ρ_2; this relation reduces to Paterson's form for head-on incidence.

If we equate the directions of flow of the two gases, one of which has been deflected by passage through the incident and reflected shock fronts, and the other deflected through the transmitted shock front, then we get an equation of the twelfth degree giving $\tan\tau$ for a given direction τ' of that final flow. Taub considered that the only real solutions were those that would approach those known for the acoustic case as $\Pi_1 \to 1$. Many of the other solutions were excluded on other grounds, for example the flow must be steady and, therefore, supersonic everywhere in a frame of reference moving with the line of intersection of the shock fronts. An interesting result was that total reflection cannot occur for incident waves of non-zero strength.

It follows from momentum considerations that there must always be a transmitted shock. If the second fluid has a relatively high acoustic impedance, we should expect the same kind of phenomena as for reflection at a rigid wall, namely regular reflection for a range of angles of incidence including head-on, and a modification of Mach reflection for other angles. On the other hand, if the second medium has a relatively low acoustic impedance, we should expect much the same phenomena as for reflection at a free surface, namely the regular type giving a reflected rarefaction wave intersecting the incident wave in the free surface, and also the second type in which a weaker shock reaches the surface. Taub has found by numerical calculation that in addition there is in some cases a narrow range of angles of incidence for which a reflected shock is possible.

(vi) *Diffraction Problems*

The earliest work on this subject was carried out by Sommerfeld (1895), but his results are exact only for sound pulses.

The diffraction of a sound pulse by a semi-infinite screen, normal to the direction of propagation of the pulse and bounded by a straight edge, has been treated by Friedlander (1946). As with Sommerfeld, the problem was simplified by assuming that all disturbances travelled with the velocity of sound in undisturbed air. Hence the mathematics could be based on a velocity potential satisfying the wave equation. The analysis was reduced to finding a suitable potential function to conform with conditions on a known boundary. Friedlander worked out in detail the case of an incident pulse with a discontinuous front, followed by a decay, of the form

$$\delta p = p'(1-t)e^{-t},$$

where p' is a small and constant pressure. This pulse includes a suction phase, and therefore gives a fair approximation to the pressure wave caused by an explosion. Other cases were considered, for instance the diffraction problem of a plane sound pulse approaching the apex of a wedge. A very interesting discussion was given of the case where the angle of the wedge was arbitrarily small. The object here was to elucidate the paradox that a sound pulse incident at nearly grazing incidence on an infinite rigid plane wall gives a doubling of the over-pressure, whereas, if the pulse runs parallel with the wall, the over-pressure is simply that in the undisturbed pulse. The diffracted pulse from the edge of the wedge behaves in such a way that the apparent contradiction is reconciled.

Friedlander's work has been extended by Fox (1948), who gave some theorems on diffraction by screens of any shape and then calculated the diffraction of a pulse consisting of an instantaneous pressure rise followed by a constant pressure sustained for an infinite time. Owing to the linearity of the wave equation, the effect of any pressure pulse can be calculated by the superposition of a number of positive and negative pulses of this type arriving at different times. Fox made use of Huyghens' principle as given by Kirchhoff, in which the pressure at any point Q can be expressed as an integral over a surface S surrounding Q, the contribution from any element dS distant r from Q being calculated at the earlier time $t-r/c_0$ at which a pulse would have had to leave dS in order to arrive at Q at time t. For S, he chose the plane of the screen together with an infinitely large hemisphere behind it, contributing nothing to the integral. He then showed that the pressure at Q could be obtained either by integrating p over the back of the screen or by integrating $\psi \equiv \partial p/\partial x$ over the apertures. By choosing Q to be a point in an aperture, an integral equation for ψ was obtained, and this was simplified by means of a Laplace transformation.

The case of a strip was treated by successive approximations. Taking the strip to be infinitely wide, Friedlander's solution ψ_0 for ψ over the semi-infinite aperture was obtained. Fox then calculated ψ_1, the change in ψ due to a distribution ψ_0 over a second semi-infinite aperture beyond the other edge of the strip; then the effect in the first aperture of a distribution ψ_1 over the second aperture was calculated, and so on, by successive approximations. Alternatively a series of approximations to p over the back of the strip could be calculated by a similar process. Since ψ was zero over the back of the screen and p was equal to the incident pulse over any aperture, the solution could then be completed.

A different approach has been made by Lighthill (1949), who transformed the problem to a linear one, not by taking constant velocity of propagation, i.e. very weak shocks, but by taking very small angles of deflection so that the disturbance of the flow was small. He took a semi-infinite plane shock, with constant conditions everywhere behind the front. The shock travelled parallel to a plane wall until at some instant it passed a line beyond which the direction of the wall was slightly changed. Then either diffraction or reflection occurred, depending on whether the corner in the wall was convex or concave.

Just as for reflection, the diffraction problem is a quasi-steady one, the disturbed region expanding steadily with time. Lighthill eliminated the time by choosing as independent variables $X=(x-w_1t)/c_1t$ and $Y=y/c_1t$, where w_1 and c_1 are fluid and sound velocities behind the incident shock and outside the disturbed zone, x is measured parallel to the initial wall direction, and the origin of x and y is at the corner. The fluid velocity components were divided by w_1 to obtain the dimensionless quantities W and V. Since the disturbance is small, W is near unity, while V is of the first order of small quantities, and to that order contributes nothing to the shock energy or pressure. The shock front therefore remains nearly plane, its direction at any point being given by

$$dX/dY = -V/W.$$

For $w_1<c_1$, i.e. subsonic flow behind the shock, Lighthill found that the boundary of the diffracted region is the shock front and a portion of a cylinder of unit radius and centre the origin in the (X,Y) plane. For strong shocks, i.e. $w_1>c_1$, the corner lies outside the cylinder. Part of the boundary is formed by the tangent plane from the corner to the cylinder. A Prandtl–Meyer expansion,

confined to a wedge with edge at the corner, then produces a uniform flow parallel to the new direction of the wall inside the region bounded by the tangent plane, the cylinder and the wall. Figures 5(*a*) and (*b*) show the boundaries in the two cases.

Lighthill had still to solve the problem inside the cylindrical region. Since the disturbances are weak, he was able to linearize the equation of motion by using as coefficients the constant values of w_1, ρ_1 etc. in the undiffracted region behind the shock front. His equations then reduced to the same form as those derived by Busemann (1943) for the 'cone field' problem. In this problem the flow is everywhere steady, supersonic and isentropic, while fluid velocities are constant along all straight lines through the origin. Busemann linearized his problem by assuming the departures from uniform flow to be small. Following Busemann, Lighthill used a transformation which reduced the partial differential equation for the pressure to Laplace's equation. The only problem remaining was then to find solutions to fit the transformed boundary conditions.

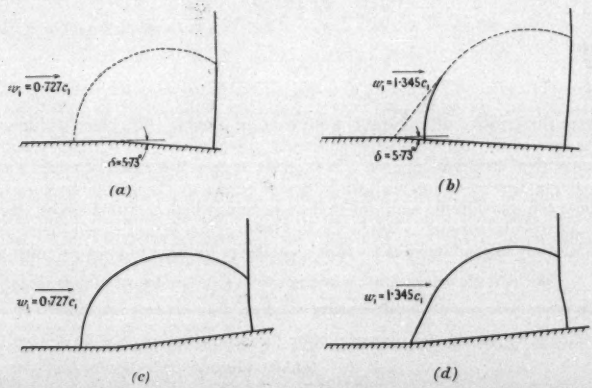


Figure 5. These diagrams illustrate Lighthill's theory of the diffraction patterns resulting when a plane shock wave moves parallel to a rigid wall which at some stage slightly changes its direction. (*a*) and (*c*) are subsonic cases (i.e. the fluid velocity behind the shock relative to the point where the wall changes direction is less than the velocity of sound behind the shock); (*b*) and (*d*) are supersonic cases.

Lighthill calculated the pressure on the wall at various points and found that in the subsonic case there was a very narrow low pressure region at the corner, followed by a rise to a maximum about midway between the corner and the shock front, and then a slight fall on approaching the front. In the supersonic case, the pressure on the wall fell abruptly at the corner, and was then uniform up to the cylindrical region, inside which it rose steadily up to the shock front.

In Figures 5(*c*) and (*d*) the boundaries are drawn for a concave corner giving Mach reflection of the shock. The cylindrical boundary is now a shock front and so has a slightly larger radius and may not be exactly circular. Exact solutions were not worked out in these cases, but the diagrams are qualitatively correct and are shown for comparison with the diffracted cases.

§4. SIMILARITY SOLUTIONS OF ONE INDEPENDENT VARIABLE

In the sections on reflection, refraction and diffraction we stated that if the process started at some finite time, e.g. when a shock front struck the edge of a wedge, then the resultant flow pattern, having no dimension imposed on it, could be one which expanded steadily with time. By changing our independent variables to x/t, y/t we can reduce the problem to one involving only two independent variables. The shape of the curved boundary surface, or surfaces, must be determined, while the flow outside is, of course, undisturbed. A major difficulty arises from the fact that the shape of the boundary is often not obvious, and where this is so the shape may be difficult to calculate. So far no exact solutions have been obtained.

A few cases have arisen in which there is only one space variable, and by changing to x/t as independent variable the equations are immediately reduced to ordinary differential equations which can be integrated numerically. The equations of continuity and motion under isentropic conditions are

$$\frac{\partial \rho}{\partial t} + u\frac{\partial \rho}{\partial x} + \rho\frac{\partial u}{\partial x} = -\frac{\kappa \rho u}{x} \qquad \ldots\ldots (25)$$

and

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = -\frac{c^2}{\rho}\frac{\partial \rho}{\partial x}, \qquad \ldots\ldots (26)$$

where the velocity of sound c is a function of ρ only. For plane disturbances κ is zero; for cylindrically symmetrical disturbances $\kappa = 1$; for spherically symmetrical disturbances $\kappa = 2$. We can see at once that a pair of values ρ and u satisfying these equations also satisfy them if x and t are changed in the same ratio. There can therefore be solutions of the equations which are functions of x/t only.

A particular example, given by Taylor (1946), is that of the air wave surrounding a uniformly expanding sphere. For a similarity solution to exist we must have

$$\left\{\frac{\partial}{\partial t} + \frac{x}{t}\frac{\partial}{\partial x}\right\}(u,\ p \text{ or } \rho) = 0. \qquad \ldots\ldots (27)$$

Writing y for x/t, the equation of motion for isentropic flow may be written

$$(u-y)\frac{du}{dy} = -\frac{c^2}{\rho}\frac{d\rho}{dy} \qquad \ldots\ldots (28)$$

and the equation of continuity

$$\frac{u-y}{\rho}\frac{d\rho}{dy} + \frac{du}{dy} + \frac{2u}{y} = 0. \qquad \ldots\ldots (29)$$

Now, from the equation of state,

$$\frac{1}{\rho}\frac{d\rho}{dy} = \frac{1}{\gamma - 1}\,;\ \frac{1}{c^2}\frac{dc^2}{dy}, \qquad \ldots\ldots (30)$$

and we can eliminate ρ from (28) and (29). Using the dimensionless variables

$$\left.\begin{aligned} \xi &= u/y, \\ \eta &= c^2/y^2, \\ \text{and} \quad z &= \ln y \end{aligned}\right\} \qquad \ldots\ldots (31)$$

we can obtain the two ordinary differential equations

$$\frac{d\eta}{d\xi} = \frac{2\eta}{\xi} \cdot \frac{\eta + (\gamma+1)\xi - \gamma\xi^2 - 1}{3\eta - (1-\xi)^2} \qquad \ldots\ldots (32)$$

and

$$\frac{dz}{d\xi} = -\frac{1}{\xi} \cdot \frac{\eta - (1-\xi)^2}{3\eta - (1-\xi)^2}. \qquad \ldots\ldots (33)$$

The first of these two equations contains ξ and η only, and Taylor solved it numerically for a series of initial values of η defined by $\eta_2 = 1/\alpha^2$ for $\xi = 1$. The meaning of α is that it is the ratio of the velocity of expansion of the sphere to the local velocity of sound in the air near its surface. Taylor plotted a series of curves, which we shall call α-curves, giving η as a function of ξ for the given values of α. He found, as is evident from the differential equation, that as ξ diminishes from unity, η first falls to a minimum for which the ξ value is greater than 0, and then rises to pass vertically through the point (0, 1). By examining (33) we can see that z increases at first and reaches a maximum at a point on the curve

$$\eta = (1-\xi)^2 \qquad \ldots\ldots (34)$$

which is higher than the curve

$$\eta = (1-\xi)^2 - (\gamma-1)\xi(1-\xi) \qquad \ldots\ldots (35)$$

which intersects all α-curves in points where $d\eta/d\xi = 0$. The curve (34) starts from the point (1, 0), i.e. below the point $(1, 1/\alpha^2)$ from which the α-curve starts, and therefore crosses it before the latter reaches its minimum. There is therefore a maximum value of z at some value of $\xi > 0$ on each α-curve. The solution cannot be valid for smaller values of ξ because it is impossible that there shall be two values of ξ for the same value of z, i.e. two fluid velocities at the same point. There cannot therefore be any continuous transition from conditions near the surface of the sphere to undisturbed air. Taylor's solutions are such that the front of the disturbance is always a shock front.

Inside the disturbed region Taylor found for the pressure distribution

$$p/p_2 = (\eta x^2 \alpha^2 / X^2)^{\gamma/(\gamma-1)}, \qquad \ldots\ldots (36)$$

where p_2 is the absolute pressure at the surface of the sphere whose radius is X. For the fluid velocity distribution he found

$$u = \xi \frac{x}{X} \frac{dX}{dt}.$$

It is interesting to note that the velocity distribution for $\alpha = 0{\cdot}2$ differed noticeably from that calculated from the acoustic approximation although the pressure discontinuity in the shock front was only about 10^{-19} atmospheres!

Another similarity solution given by Taylor (1941) is that for a blast wave produced by a very intense explosion in a small volume. For the expanding sphere problem considered above the shock wave energy increased as the cube of its radius, but in this case the mechanical energy in the wave is decreasing. By writing

$$p/p_0 = y = f_1/X^3, \quad \rho/\rho_0 = \psi, \quad u = \phi_1/X^{3/2}, \qquad \ldots\ldots (37)$$

where X is the radius of the shock front and f_1, ψ and ϕ_1 are functions of $\eta = x/X$ only, it can be shown that these similarity relationships are consistent with the equations of motion and continuity for a perfect gas. The equation of motion

can be satisfied if $dX/dt = A/X^{3/2}$ where A is a constant. Making the further substitutions

$$f = f_1 c_0^2/A^2; \qquad \phi = \phi_1/A \qquad \ldots\ldots (38)$$

three ordinary simultaneous differential equations are obtained, namely

$$\left.\begin{aligned} (\eta-\phi)\phi' - f'/\gamma\psi + 3\phi/2 &= 0, \\ \psi'/\psi - (\phi' + 2\phi/\eta)/(\eta-\phi) &= 0, \\ 3f + \eta f' + \gamma\psi'(\phi-\eta)f/\psi - \phi f' &= 0, \end{aligned}\right\} \qquad \ldots\ldots (39)$$

where the dashes denote differentiation with respect to η. These equations contain only one parameter, namely γ. They can be rearranged to give each differential coefficient as a function of the undifferentiated variables and then integrated. The boundary conditions at the shock front have been taken to be the limiting ones for very strong shocks, namely

$$\left.\begin{aligned} \rho_1/\rho_0 &= (\gamma+1)/(\gamma-1), \\ u_1/U &= 2/(\gamma+1), \\ U^2/c_0^2 &= (\gamma+1)y_1/2\gamma, \end{aligned}\right\} \qquad \ldots\ldots (40)$$

from which corresponding initial conditions for f, ψ and ϕ follow.

Taylor integrated the equations step by step for $\gamma = 1{\cdot}4$ and found that at any instant the fluid velocity increased almost in proportion to the radius right out to the shock front. Going inwards from the shock front, the fluid density decreased rapidly to 1/10, 1/100 and 1/1,000 of its maximum value for $\eta = 0{\cdot}84$, $0{\cdot}64$ and $0{\cdot}49$, while the pressure fell to 30% of its maximum value at $\eta = 0{\cdot}75$ and then remained almost constant. The temperature therefore rose to enormous values near $\eta = 0$. Taylor also found that nearly 45% of the shock wave energy was dissipated in heating the air by the time that the shock front pressure had fallen to 20 atmospheres. An approximate calculation for $\gamma = 5/3$ gave very similar results.

Calculations on the same lines for $\gamma = 1{\cdot}2$ and $1{\cdot}1$ have been carried out by Hirschfelder (1946). In these cases the pressure and density gradients behind the shock front at any instant are steeper, but the uniform pressure near the centre is about 40% and 50% of its value at the front in the two cases. The energy dissipations by the time the shock pressure has fallen to 20 atmospheres are 60% and 80% respectively.

Another approach to the same problem has been made by Kynch (1942), who used as dependent variables the Lagrangian coordinate x giving the original position of a particle, in addition to the Eulerian coordinate X giving its position at some later time t. He used for independent parameter the ratio θ of the kinetic energy of an element of fluid to its internal energy

$$\theta = \gamma(\gamma-1)u^2/2c^2. \qquad \ldots\ldots (41)$$

This is a convenient parameter, since for strong shocks θ is very nearly unity at the shock front and almost certainly decreases to zero at the origin. By taking an inner sphere of fluid and equating its rate of loss of energy to its rate of doing work on the fluid outside it, Kynch was able to deduce the following equation for the velocity of any fluid element:

$$\left[\frac{\partial X}{\partial t}\right]_x = \frac{dY}{dt}\,\frac{X}{Y}\,\frac{1+u^2/2E}{\gamma+u^2/2E}, \qquad \ldots\ldots (42)$$

where Y is the radius of the shock front at any instant. Since the right-hand side of equation (42) is very insensitive to changes in u^2/E, the reason why Taylor found the fluid velocity at any instant to be nearly a linear function of the radius is revealed.

Kynch integrated his equations and obtained closed expressions which could be evaluated. Good agreement in the case $\gamma = 5/3$ was found with Taylor's approximate solution. It is interesting to note that all solutions of this kind give a shock front radius increasing in proportion to $t^{2/5}$.

These solutions closely resemble those for conditions inside a spherical detonation wave, the main differences being in the outer boundary condition and in the equation of state. The temperature at the centre, however, is finite. Solutions have been given by Taylor (1941) and by Dasgupta and Penney (1943), and both are described by Cole (1948).

Finally, mention must be made of two other types of fluid flow which differ from those previously considered in that usually they describe conditions in only part of the field. The first is the Prandtl–Meyer flow mentioned earlier and of which a description is given by Taylor and Maccoll (1935). This flow is almost invariably an expansion since compression of a supersonic flow usually leads to the formation of a shock.

Taylor and Maccoll (1935) describe the cases of a wedge or a cone placed in air flowing at supersonic speed in a direction parallel to the line of symmetry of the body. In the first case the direction of flow is changed to flow parallel to the wedge surface through the mechanism of an oblique plane shock front passing through the leading edge of the wedge. The flow between the shock front and the wedge is uniform. In the second case cylindrical divergence causes the flow to be non-uniform behind the conical shock front, but for a wide range of conditions numerical solutions, which agreed with experiment, were obtained by assuming that conditions were uniform over any conical surface with the same axis and apex as the conical body. This flow is like a Prandtl–Meyer flow in its dependence on one angle only, but differs from it in that the fluid velocity is often subsonic in all or part of the field, so that the condition that the velocity component perpendicular to the radius vector at any point should be sonic is not fulfilled. It was found that under some conditions, e.g. too large a cone angle or too low an air velocity, no solutions of the type sought were possible, and experiment showed that the shock wave was then detached from the apex of the cone and remained stationary a short distance in front of it. Some excellent photographs of such flows are given by Maccoll (1937), and these also show a three-dimensional modification of a Prandtl–Meyer expansion round the base of the cone.

§5. NON-SIMILARITY SOLUTIONS

(i) *A Generalization of Riemann's Characteristics*

Mention was made in the opening paragraphs of this Report of Riemann's treatment of one-dimensional pulses in a fluid, there being no entropy changes caused by the pulses. The following generalization of Riemann's analysis to cases where there are entropy gradients and the pulses are in one dimension, or are spreading symmetrically in two or three dimensions, retains the concept of two sets of waves moving in opposite directions, and regards the actual motion as an interference pattern produced by the two.

We assume that conditions are uniform over any plane, cylindrical or spherical surface at distance x from a fixed plane, line or point as the case may be. Let p be the pressure, ρ the density, u the mass velocity and S the entropy at the field point. The equations of continuity and motion, as usual, are

$$\frac{\partial \rho}{\partial t} + u\frac{\partial \rho}{\partial x} + \rho\frac{\partial u}{\partial x} = -\frac{\kappa \rho u}{x}, \qquad \ldots\ldots (43)$$

$$\rho\frac{\partial u}{\partial t} + \rho u\frac{\partial u}{\partial x} + \frac{\partial p}{\partial x} = 0, \qquad \ldots\ldots (44)$$

where $\kappa = 0$, 1 or 2 for plane, cylindrical or spherical flow respectively.

If the flow is isentropic, then p is a function of ρ only, and either p or ρ may be eliminated from (43) and (44).

If the flow is not isentropic, we choose to regard ρ and S as the independent thermodynamic variables and write

$$p = F(\rho, S), \qquad \ldots\ldots (45)$$

where F is a function known from the thermodynamic properties of the fluid.

The approximation is made that dissipative mechanisms such as thermal conductivity, viscosity, radiation, etc. may be neglected in the pulse. The entropy is therefore constant following the motion of any element of the fluid. Hence

$$DS/Dt = 0. \qquad \ldots\ldots (46)$$

The generalized Riemann function may be defined

$$\sigma(\rho, S) = \int_{p_1}^{p} \sqrt{(\partial\rho/\partial p)}\, dp/\rho, \qquad \ldots\ldots (47)$$

the integral being taken along a path of constant entropy. The pressure p_1 is some convenient pressure, either atmospheric or zero.

Immediate consequences from (47) are

$$\rho c(\partial\sigma/\partial p)_S = 1, \quad \rho(\partial\sigma/\partial\rho)_S = c. \qquad \ldots\ldots (48)$$

For a fluid whose equation of state is of the form (14) one finds

$$\sigma = 2(c - c_1)/(n-1). \qquad \ldots\ldots (49)$$

A convenient value for p_1 in this case is zero.

Perfect gases are included in this class as special cases; for these $B(S)$ is zero and n is γ. Consequently, for a perfect gas, we may take

$$\sigma = 2c/(\gamma - 1) \qquad \ldots\ldots (50)$$

if the standard state is taken to be one where $p_1 = 0$, or

$$\sigma = 2(c - c_0)/(\gamma - 1) \qquad \ldots\ldots (51)$$

if the standard state is taken to be atmospheric pressure.

We define two characteristic Riemann functions r and s by

$$r = \sigma + u; \quad s = \sigma - u. \qquad \ldots\ldots (52)$$

Now
$$dr = \frac{\partial r}{\partial x}dx + \frac{\partial r}{\partial t}dt$$
$$= \left\{\frac{\partial\sigma}{\partial\rho}\frac{\partial\rho}{\partial x} + \frac{\partial\sigma}{\partial S}\frac{\partial S}{\partial x} + \frac{\partial u}{\partial x}\right\}dx + \left\{\frac{\partial\sigma}{\partial\rho}\frac{\partial\rho}{\partial t} + \frac{\partial\sigma}{\partial S}\frac{\partial S}{\partial t} + \frac{\partial u}{\partial t}\right\}dt.$$

The $\partial u/\partial t$ term may be eliminated by using the equation of motion (44), and the $\partial p/\partial t$ term may be eliminated by using the equation of continuity (43). Other terms cancel in view of (48). It will be found that if dx is taken equal to $(u+c)\,dt$, then several more terms cancel, and there remains the relationship

$$dr = dt\left[-\frac{\kappa uc}{x} + c\left(\frac{\partial\sigma}{\partial S}\right)_p\left(\frac{\partial S}{\partial x}\right)_t\right]. \qquad \ldots\ldots (53)$$

This is the generalization of the Riemann result that for a plane wave in an isentropic fluid a point moving with local speed $(u+c)$ carries the r value unchanged.

Similarly, it will be found that

$$ds = dt\left[-\frac{\kappa uc}{x} - c\left(\frac{\partial\sigma}{\partial S}\right)_p\left(\frac{\partial S}{\partial x}\right)_t\right] \qquad \ldots\ldots (54)$$

along the line $dx=(u-c)\,at$. The lines $dx=(u\pm c)\,dt$ are called characteristic lines. To distinguish them we shall call them r and s lines respectively.

The term $-\kappa uc/x$ in dr and ds represents the local change in the Riemann characteristic due to the cylindrical or spherical expansion ($\kappa=1$ or 2), while the term is absent for plane waves ($\kappa=0$). The second terms in dr and ds represent the local changes in the characteristics, due to the entropy gradient.

Equations (52), (53) and (54) contain, in principle, all that is required to determine the development of the motion resulting from any initial disturbance in one, two or three dimensions. Knowing r, s and S throughout the initial disturbance, the subsequent changes in these functions can be calculated step by step, and thus the complete solution followed as far as desired, with any accuracy desired, the only limitation being the computational effort required.

Penney and Dasgupta (1942) have used the Riemann characteristics to follow the development of the blast wave caused by a spherical explosion in water and in air. The entropy gradient term was omitted for the wave in water, but was retained for the wave in air. Another problem which has been considered is the pulse caused by the sudden starting to constant velocity of a piston in a cylinder containing air at N.T.P., $\gamma=1{\cdot}4$, followed by a sudden stoppage of the piston. The motion has been carried to a stage where the length of the pulse is 20 times the length of the pulse at the instant the piston stopped. Two cases, corresponding to 100 atmospheres and 10 atmospheres pressure in the initial high pressure region, were taken.

In the numerical work just described about 20 entries for the space coordinate were taken. The time interval dt was chosen so that the r-characteristic moved about one space position outwards and the s-characteristic one space position inwards. The local changes dr and ds throughout each step were calculated from the local values of c, u, S and x_n, and the new positions of the r and s points determined from the local $(u\pm c)$ values. The new r and s values on the x_n grid were then found by linear interpolation. All of this work could be done in tabular form. The position of the shock was always kept exactly, and conditions there were fitted by a separate calculation. About four to six hours' work was needed per step, and the only equipment used was a hand calculating machine. With steps of this size errors could be kept at 1% or less per step, but after 30 steps accuracy was beginning to be lost. The only sensitive check that could be applied was conservation of energy. In the case of the spherical blast wave in water the energy showed very good agreement with the initial value after a six-fold expansion

in the pulse, but in the case of the air-blast problem and the piston problem the energy was 15–20% too large after a six-fold expansion in the pulse. Problems of this type will no doubt be solved much more accurately with electronic computing machines; it is unlikely that the Riemann characteristics will be used, as it will probably prove simpler to deal directly with the original hydrodynamical equations in Lagrangian form, or some other suitable coordinate system.

We now justify an assumption which has previously been made by all authors considering the spherical expansion of intense pulses through water. The assumption to which we refer is that the entropy gradient terms may be neglected in comparison with the spherical expansion terms. Referring to (53), it will be seen that the ratio of the two terms, using $\kappa = 2$, is

$$\frac{2u}{x} : \left(\frac{\partial \sigma}{\partial S}\right)_p \left(\frac{\partial S}{\partial x}\right)_t.$$

Now in order to find the required differential coefficient of σ one should use the definition of σ given in (47) and evaluate a two-dimensional table of values with S and p as independent variables, based on Bridgman's data. Such an elaborate task is unnecessary here as we are concerned only with orders of magnitude. Accordingly we use the formula (49), which is admittedly based on a slightly inaccurate equation of state, and form the required differential coefficient. Since n for water is approximately 7, we have

$$\left(\frac{\partial \sigma}{\partial S}\right)_p = \frac{T}{3}\left\{\left(\frac{\partial c}{\partial T}\right)_p - \left(\frac{\partial c_1}{\partial T}\right)_p\right\}, \qquad \ldots\ldots (55)$$

where the temperature T and the pressure p are the values just behind the shock front, c is the velocity of sound just behind the shock front, and c_1 is the velocity of sound in the water just behind the shock front when it has returned adiabatically to atmospheric, or zero, pressure.

Let us take the extreme case of the pressure wave just leaving the surface of an explosive detonated under water. The shock front pressure is of the order 50,000 atmospheres. From Table 1 the water will be about 150° C. hotter than it was initially, and when the pressure has returned to zero it will still be 50° C. hotter than it was initially. Experimentally, it is known that the velocity of sound in water at atmospheric pressure increases by about 3 metres per second per degree Centigrade, but that this rate falls to zero at 50° C., and at higher temperatures becomes slightly negative. With adequate approximation for our purpose we may therefore take $(\partial c_1/\partial T)_p$ in equation (55) as 3 m/sec/° C.

Bridgman's data on water show that the velocity of sound at high pressures in water at constant pressure changes but little with temperature, the actual change being negative and of order 0·1 m/sec/° C. We therefore neglect the term $(\partial c/\partial T)_p$ in equation (55).

From Table 1 the entropy change at a strong shock front in water is proportional to the shock pressure, and for a shock of 50,000 atmospheres is 0·17 calories per degree per gramme. It is known that the shock pressure in the pulse falls off approximately as the inverse of the radius. Hence initially $(\partial S/\partial x)_t = -0{\cdot}17/x$. This estimate is perhaps too small by a factor 2 since peak pressure falls off faster than radius, and also because the term we require is the space gradient, whereas what we have evaluated is the rate of entropy change with shock front radius.

The value of u at 50,000 atmospheres from Table 1 is 1,250 metres per second.

Collecting together the above results, we find that the ratio of the radial expansion terms to the entropy gradient terms in the expressions for the rate of change of the Riemann characteristics is approximately $2{,}500/x : 150/x$. Thus the radial expansion terms are about 10–20 times as great as the entropy gradient terms even in the very early stages, where the entropy gradient has its greatest relative effect.

We conclude that an expanding pulse in water may be treated as isentropic provided that the initial pressures do not greatly exceed 50,000 atmospheres. At much higher pressures the entropy terms begin to count, since u increases only as $\sqrt{p}$, whereas the entropy terms increase as p. However, at these pressures the equation of state which we have used for water may not be valid.

(ii) *Transition of a Simple Expanding Wave into a Shock*

We shall now make use of the Riemann differential relations to show that if initially a cylindrical or spherical disturbance, free of shocks, is caused within a finite region, then the expanding edge of the disturbance will, sooner or later, change into a shock front, provided the pulse is such that the pressure increases as we pass inwards away from the edge. Until shock fronts form, the fluid is isentropic and the entropy gradient terms in (53) and (54) are absent.

We must calculate the rate of change with time of the steepness of the pressure (or fluid velocity) gradient at the front edge of the disturbance. At that edge $u=0$, $c=c_0$, so that it follows the r line $dx=c_0\,dt$, where c_0 is the velocity of sound in undisturbed fluid. We are required to form differential coefficients following an r line and shall use the operator

$$\frac{D_1}{Dt} \equiv \frac{\partial}{\partial t} + (u+c)\frac{\partial}{\partial x}. \qquad \ldots\ldots (56)$$

The corresponding operator for differentiation following an s line is

$$\frac{D_2}{Dt} \equiv \frac{\partial}{\partial t} + (u-c)\frac{\partial}{\partial x}. \qquad \ldots\ldots (57)$$

Then the Riemann differential relations (53) and (54) are simply

$$\frac{D_1 r}{Dt} = \frac{D_2 s}{Dt} = -\frac{\kappa u c}{x}. \qquad \ldots\ldots (58)$$

Furthermore, we have the operational identities

$$\frac{D_1}{Dt} = \frac{D_2}{Dt} + 2c\frac{\partial}{\partial x} \qquad \ldots\ldots (59)$$

and

$$\frac{D_1}{Dt}\frac{\partial}{\partial x} = \frac{\partial}{\partial x}\frac{D_1}{Dt} - \frac{\partial(u+c)}{\partial x}\frac{\partial}{\partial x}. \qquad \ldots\ldots (60)$$

Consider Figure 6, which shows the mass velocity and pressure just behind the leading edge of the disturbance when the edge is at X, the time being t. Measure a small distance z backwards from X.

Taking only the linear terms in z, we may write

$$u = Gz, \quad p = p_0(1+Hz), \qquad \ldots\ldots (61)$$

and our problem is to find how G and H vary with t or X.

At time t the value of s at X is zero, so that the value at z is simply ds, obtained from $ds = -\kappa u c\,dt/x$. Clearly we must take the average value of u, which is

$\frac{1}{2}Gz$, while for dt we must take $(z/2c)$, the 2 in the denominator entering because the relative velocity of the s point and the edge of the wave is $2c$. Hence, to second order in z,

$$s_z = -\tfrac{1}{4}\kappa Gz^2/x. \qquad \ldots\ldots (62)$$

Therefore, to first order in z,

$$\sigma = u, \quad \tau = 2\sigma = 2u. \qquad \ldots\ldots (63)$$

Again, using (61), we find to first order in z,

$$c = c_0\{1 + (\gamma - 1)Hz/2\gamma\}. \qquad \ldots\ldots (64)$$

From equation (51) we find that σ for values of $p - p_0$ given by (61) is, to first order in z,

$$\sigma = c_0 Hz/\gamma. \qquad \ldots\ldots (65)$$

A comparison of (61) and (63) with (65) leads to the conclusion that

$$\gamma G = c_0 H, \qquad \ldots\ldots (66)$$

so that the slopes of the u and p curves in Figure 6 are proportional to each other.

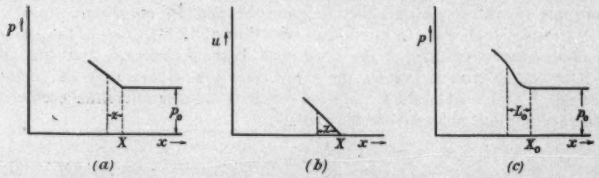


Figure 6. Diagrams needed in the proof that a pressure wave expanding symmetrically in two or three dimensions must at some stage develop a shock front somewhere within the pulse, and that a shock will ultimately lead the pulse.

Apply the operator $(D_1/Dt)(\partial/\partial x)$ of identity (60) to r in the region near the edge. We have, remembering that $\partial/\partial x = -\partial/\partial z$,

$$\frac{D_1}{Dt} G = \frac{\partial}{\partial z}\left\{-\frac{\kappa uc}{2x}\right\} + \frac{\partial(u+c)}{\partial z}\frac{\partial u}{\partial z}. \qquad \ldots\ldots (67)$$

We now perform the $\partial/\partial z$ operations on the right-hand side of (60), and proceed to the limit $z = 0$. In the limit we replace D_1/Dt by $c_0 d/dX$, since the $\partial/\partial x$ term in D_1/Dt does not appear as the r-point moves with X. Expressing the results in terms of H, rather than G, as we may do simply by using (66), we obtain

$$\frac{dH}{dX} = -\frac{\kappa}{2X} H + \frac{\gamma+1}{2\gamma} H^2. \qquad \ldots\ldots (68)$$

For the practical case of a spherical expansion ($\kappa = 2$) in air ($\gamma = 1{\cdot}4$) we have

$$dH/dX = -H/X + 6H^2/7. \qquad \ldots\ldots (69)$$

The gradient at the edge is therefore increasing only if $H > 7X/6$. The integral of (68) is

$$H = H_0 X_0/X(1 - \tfrac{6}{7} H_0 X_0 \ln X/X_0), \qquad \ldots\ldots (70)$$

where H_0 and X_0 are initial values.

The leading edge therefore becomes a shock front when

$$X = X_0 \exp(7H_0X_0/6). \qquad \ldots\ldots (71)$$

The above discussion can be generalized to show that if the expanding pressure pulse is initially like that shown in Figure 6 (*c*), L_0 being small compared with X_0, then a shock is first formed at some point behind the leading edge. The r-point that is in the front when this front first forms was initially very near to the place of greatest pressure gradient, and the radius at which the shock front forms is given by (71), where p_0H_0 is now to be interpreted as the initial pressure gradient and X_0 is the initial radius.

The complete justification for this statement is too lengthy to be given here. However, knowing that s is zero at the leading edge of the pulse, and from the differential equation for s along an s line,

$$\frac{D_2 s}{Dz} = -\frac{\kappa uc}{(c-u)x}, \qquad \ldots\ldots (72)$$

we see that s is small compared with u, and therefore compared with r, throughout the region where u increases monotonically with z as we move in from the edge of the wave, provided the thickness of that region is small compared with the radius X of the wave. This means that (63) can still be used and that the result of applying the operation (60) to r and retaining the terms of leading order gives the results stated.

In a spherical or cylindrical expanding wave, led by a pressure pulse, a suction pulse must follow. The minimum in the suction pulse must be followed by a portion of rising pressure. A shock front must ultimately form in this region, the arguments being as before. In the spherical shock waves produced by explosions this second front cannot be separated from subsequent shocks due to oscillations of the gas sphere or ground reflection, but Dumond *et al.* (1946) have found that the shock wave from the front of a projectile travelling at supersonic speed (the bow wave) has in many cases an almost perfect N form.

The above analysis was given by Penney (1941). Similar conclusions have been reached by Landau (1942).

A degenerate case of the above problem is that where the leading edge of the disturbance is quadratic in z, i.e. $u = gz^2$,

$$p = p_0(1 + hz^2). \qquad \ldots\ldots (73)$$

We now have to go to third order in differential operators and use the identity

$$\frac{D_1}{Dt}\frac{\partial^2}{\partial x^2} = \frac{\partial^2}{\partial x^2}\frac{D_1}{Dt} - 2\frac{\partial(u+c)}{\partial x}\frac{\partial^2}{\partial x^2} - \frac{\partial^2(u+c)}{\partial x^2}\frac{\partial}{\partial x}. \qquad \ldots\ldots (74)$$

This time s is of the third order in z, and r equals 2σ or $2u$ to second order. Applying (74) to r in a manner very similar to that used before, we find that the left-hand side and the first term on the right-hand side are independent of z, but that the second and third terms are linear in z. Proceeding to the limit $z = 0$, we thus have only $dg/dX = -\kappa g/2X$. The integral is

$$g = g_0(X_0/X)^{\kappa/2}. \qquad \ldots\ldots (75)$$

In this case, the front stays quadratic and the curvature decreases in proportion to $X^{-\kappa/2}$. Of course a shock must form behind the leading edge and gradually overtake it, thus once again giving a pulse led by a shock front.

Calculations using the method of characteristics (Sauer 1947) have shown that the curvature of the front of the pressure wave remains constant in the plane case and diminishes in the cylindrical or spherical case. The calculations were not taken far enough to show the development of a shock front behind the wave front.

Similar calculations were carried out by Unwin (1941). He treated the case of spherical symmetry, his initial condition being air at rest but compressed adiabatically to a density distribution given by

$$\rho - \rho_0 = (\rho_1 - \rho_0) \exp(-4r^2).$$

He put the left-hand sides of the equations of continuity and momentum in total form following the motion of a particle and then integrated numerically to follow the motion of several spherical shells of fluid. He found that the fluid over-expanded and then over-contracted, so giving a series of fairly sharp pressure pulses separated by long suction phases. Once again the numerical work was not taken sufficiently far to show the development of shocks.

The changing form of a spherically expanding pressure pulse in air has been demonstrated experimentally by Jehu (1941). A balloon, normally used for meteorological observations, was filled with a mixture of hydrogen and oxygen. A spark was triggered inside the balloon and the gas mixture exploded. The gas mixture does not instantly pass into a detonation, and by arranging the experimental conditions suitably a pressure pulse free of shocks is sent into the surrounding air. By placing a number of piezoelectric gauges at various distances, the changing form of the pulse in the air can be recorded; Figure 7 shows a set of records. The formation of a shock inside the pulse and the passage of this shock up to the leading edge of the pulse can be clearly seen.

(iii) *Calculation of the Propagation of Shock Waves caused by Explosions*

A complete solution for the propagation of the shock wave produced by a spherical explosive starts from the instant when the detonation wave reaches the surface of the charge. Provided we know the equation of state of the detonation products, then the distribution of pressure, velocity etc. in those products can be determined if we assume conditions inside the detonation wave to be quasi-steady, i.e. functions of x/X only, where X is the radius of the detonation wave. A description of the results obtained is given by Cole. Outside the surface of the explosive we have air, water or other fluid at rest. The calculation can then be carried forward using the non-isentropic Riemann equations with three sets of characteristics. For underwater explosions we have shown that we can use the isentropic equations without much loss of accuracy. A limited amount of work has been carried out on both these lines, but the calculations were very tedious (Penney and Dasgupta 1942). An important analytical treatment has been given by Bethe and Kirkwood; this work is fully described by Cole (1948) and therefore will not be repeated here.

A semi-empirical method giving quite good accuracy for the shock waves from explosions in air has been developed by Brinkley and Kirkwood (1947a). It depends on the experimental observation that the pressure–time curve measured at a fixed point has always very nearly the same shape when expressed relative to the two parameters, peak shock over-pressure and duration of the positive pulse.

When a shock wave passes through any small element of air that element moves, and in so doing performs mechanical work on the next element; in this way the

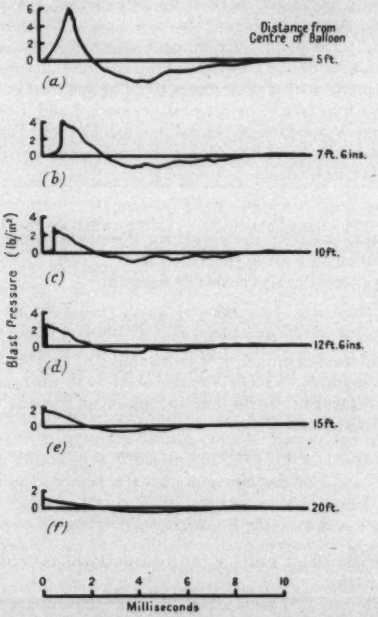


Figure 7. Experimental piezo-gauge pressure–time curves showing the gradual development of a shock in the spherically expanding pressure wave caused by the explosion of a gas-filled balloon (Jehu).

energy of the shock wave is transmitted onwards. For a spherical wave we can define the energy ϵ of the wave when its front reaches some radius R as the work done by all the elements of air which were initially on the spherical surface of that radius, i.e.

$$\epsilon(R) = \int_0^\infty 4\pi r^2 p u \, dt, \qquad \ldots\ldots (76)$$

where r is the radius of that spherical surface at any later time t as it expands with velocity u, and p is the excess of the pressure in it above atmospheric*.

The energy ϵ is a steadily decreasing function of R because the shock compression is irreversible, so that after the shock wave has passed and the elements of air have returned to atmospheric pressure they have a higher temperature than formerly. It is assumed that all the energy of a shock wave

* To obtain the total work done we should use not p but the total pressure. The extra work comes, however, from the permanent expansion of the gases behind the shock, and is allowed for later by subtracting from ϵ the increase in total heat instead of the increase in internal energy of the elements of air.

in free air is ultimately expended in increasing the total heat of the air through which it has passed.

It should be pointed out that we are here dealing with the pressure–time curve for a given element of air, and this is not quite the same as that measured by a gauge at a fixed point. Experiment shows that the shock waves from orthodox explosions in free air always give pressure–time curves of the same general shape at a fixed point, so that such curves could be specified in terms of two parameters, which we may take to be the peak pressure P and the initial slope of the pressure–time curve immediately behind it. Brinkley and Kirkwood postulated that much the same thing was true for the pressure–time curve for a given element of air or, more precisely, that

$$\epsilon(R) = R^3 f(P)/(-Dp/Dt), \qquad \ldots\ldots (77)$$

where $f(P)$ is calculable from the properties of shock waves in air and D implies differentiation following a given element of air.

These shock properties also enable us to write

$$-\frac{d\epsilon}{dR} = 4\pi R^2 h(P), \qquad \ldots\ldots (78)$$

where h is the increase in enthalpy of unit volume of air after being compressed in a shock of strength P. Finally, the rate of decay of the peak pressure P with time is largely dependent on the rate of fall-off of pressure behind the peak. Brinkley and Kirkwood show that

$$\frac{d(\log PR)}{d(\log R)} = -Rl(P)\frac{Dp}{Dt} - \eta(P),$$

which equals

$$-\frac{R^3 l(P) f(P)}{\epsilon(R)} - \eta(P) \qquad \ldots\ldots (79)$$

from (77). The functions l and η can be calculated in terms of P, and η is relatively unimportant.

Equations (78) and (79) form a pair of linear simultaneous equations which can be integrated given suitable starting conditions e.g., $\epsilon = \epsilon_1$ and $R = R_1$ for $P = P_1$. We see that if we increase ϵ_1 and R_1[3] in the same proportion while leaving P_1 unchanged the solution of the equations is unchanged. If we take our starting conditions at the surface of a spherical charge, we can make the allowed change by increasing the size of the charge, and we then have the ordinary scale factor for distance $d \propto W^{1/3}$.

In consequence of the R^2 or R^3 factor on the right-hand sides of (78) and (79), both ϵ and PR are very nearly constant when R is small. When R is very large the decreases in the functions of P more than compensate the increase in the R factors, so that we again get PR and ϵ nearly constant; the wave is then propagated very much like a sound wave. For intermediate values of R both P and ϵ lie on two-parameter families of curves.

The initial value of P_1 can be calculated from the Riemann equations. The total work done by the explosive gases in adiabatic expansion up to the time when they finally come to rest at atmospheric pressure is not all passed into the first shock wave, but the deficit is roughly compensated by additional energy derived from further combustion of the explosive gases as they mix with air.

In a subsequent paper Brinkley and Kirkwood (1947 b) have modified their theory to meet the case of the approximately conical shock wave produced by

a long cylinder of explosive detonated from one end. Their methods can be equally applied to explosions in water or any other fluid of known equation of state. They can also be applied to determine the behaviour of a shock wave generated in any other way provided that initially the pressure–time curve at some fixed point has the right shape.

(iv) *Physiological Effects of Blast*

So far in this Report no mention has been made of the destructive effect of a pressure pulse passing through a medium on the medium itself or on some structure immersed in the medium. Damage considerations are beyond our present scope, but an exception might be made in favour of the physiological effects of blast. Some outstanding work in this field was done during the recent war by Zuckerman and his collaborators (1940, 1941, 1942). Not all of this work has so far been published, but we understand from Professor Zuckerman that in due course most of it will be.

The human body is remarkably resistant to damage by a blast wave in air. Ear drums are by far the most vulnerable point, and these burst when the peak pressure in the blast is from 5 to 15 lb/in^2, depending on the direction of the head with respect to the blast and on the individual. At very much higher peak pressures, of the order 200–250 lb/in^2, general damage of a serious type is caused. Haemorrhage is caused in the lungs and at other places where tissue has an internal boundary surface with an air or gas cavity. Damage of the same type is caused to men or animals swimming on the surface, or under the surface, from underwater explosions, and the critical peak pressure is of the order one ton to the square inch. Small animals are more vulnerable than large ones to air blast. Mice succumb from lung haemorrhages at about 25 lb/in^2 shock over-pressure. The duration of the blast appears to be relatively unimportant provided it is short.

Zuckerman's results can be qualitatively understood by supposing that the hydrostatic pressure in the blast accelerates the chest walls inward, the acceleration being inversely proportional to the mass per unit area of the chest wall (i.e. roughly inversely proportional to the cube root of the body weight). The chest wall acts as a piston and projects a pressure pulse into the lung. Pressure gradients in this pulse can cause the lung tissue to tear. A first attempt at a mathematical theory relating to these observations has been given by Penney and Bickley (1943). The outstanding hydrodynamical problem is the solution of the propagation of finite pulses through a material like lung tissue, cork, sorbo-rubber or foam.

ACKNOWLEDGMENTS

Reference has been made in this Report to a number of official papers written during the war; some of these papers have only recently been declassified and, of these, some may never be published in the scientific journals. The authors wish to express their thanks to Dr. M. Blackman, Mr. V. J. Jehu, Mr. J. J. McEnhill, Dr. G. J. Kynch, Dr. W. G. Marley, Sir Geoffrey Taylor and Prof. S. Zuckerman for permission to quote from their declassified work before publication, and to Dr. M. J. Lighthill for permission to reproduce Figure 5 from his paper. Thanks are also expressed to the Chief Scientific Adviser, Ministry of Home Security, the Chief of the Royal Naval Scientific Service, and the Chief Scientist, Ministry of Supply, for permission to quote freely from reports written for their Departments.

REFERENCES

BECKER, R., 1922, *Z. Phys.*, **8**, 321.

BRINKLEY, S. R., and KIRKWOOD, J. G., 1947 a, *Phys. Rev.*, **71**, 606 ; 1947 b, *Ibid.*, **72**, 1109.

BRINKLEY, S. R., KIRKWOOD, J. G., and RICHARDSON, J. M., 1944, *O.S.R.D. declassified report* 3550 ; 1945, *O.S.R.D. declassified report* 5481.

BUSEMANN, A., 1943, *Luftfahrtforschung*, **20**, 105.

COLE, R. H., 1948, *Underwater Explosions* (New Jersey : Princeton University Press).

COWLING, T. G., 1942, *Phil. Mag.*, **33**, 61.

COURANT, R., and FRIEDRICHS, K. O., 1948, *Supersonic Flow and Shock Waves* (New York : Interscience Publishers).

DASGUPTA, H. K., and PENNEY, W. G., 1943, *Ministry of Home Security report* (unpublished).

DUMOND, J. W. M., COHEN, E. R., PANOFSKY, W. K. H., and DEEDS, E., 1946, *J. Acoust. Soc. Amer.*, **18**, 97.

EKMAN, V. W., 1908, *Publications de Circonstance No.* 43 (Conseil Permanent Internationale pour l'Exploration de la Mer).

FOX, E. N., 1948, *Phil. Trans. Roy. Soc.* A, **241**, 71.

FRIEDLANDER, F. G., 1946, *Proc. Roy. Soc.* A, **186**, 322.

FUCHS, K., KYNCH, G. J., and PEIERLS, R. E., 1946, *Ministry of Supply and Los Alamos declassified report* LADC. No. 319.

HIRSCHFELDER, J. O., 1946, *Los Alamos declassified report* 74.

HIRSCHFELDER, J. O., and MAGEE, J. L., 1946, *Los Alamos declassified report* 122.

JEHU, V. J., 1941, *Ministry of Home Security report* (unpublished).

KYNCH, G. J., 1942, *Ministry of Supply report* (unpublished).

LANDAU, L. D., 1942, *J. Phys. Acad. Sci. U.S.S.R.*, **6**, 229.

LEWIS, B., and VON ELBE, G., 1938, *Combustion, Flames and Explosions of Gases* (Cambridge : University Press).

LIGHTHILL, M. J., 1949, *Proc. Roy. Soc.* A, **198**, 454.

MACCOLL, J. W., 1937, *Proc. Roy. Soc.* A, **159**, 459.

MARLEY, W. G., 1942, *Ministry of Supply report* (unpublished).

VON NEUMANN, J., 1943, *U.S. Navy Dept., Bureau of Ordnance declassified Explosive Research Report* 12.

NOVIKOV, I. I., 1948, *C.R. Acad. Sci. U.S.S.R.*, **59**, 1545.

PATERSON, S., 1948, *Proc. Phys. Soc.*, **61**, 119.

PENNEY, W. G., 1941, *Ministry of Home Security reports* (unpublished).

PENNEY, W. G., and BICKLEY, W. G., 1943, *Ministry of Home Security report* (unpublished).

PENNEY, W. G., and DASGUPTA, H. K., 1942, *Ministry of Home Security report* (unpublished).

PENNEY, W. G., and DAVIES, D. R., 1941, *Ministry of Home Security report* (unpublished). (See also DAVIES, D. R., 1948, *Proc. Phys. Soc.*, **61**, 105.)

POLACHEK, H., and SEEGER, R. J., 1944, *U.S. Navy Dept., Bureau of Ordnance declassified Explosives Research Reports* 13 and 14.

RICHARDSON, J. M., ARONS, A. B., and HALVERSON, R. R., 1947, *J. Chem. Phys.*, **15**, 785.

SACHS, R. G., 1946, *Phys. Rev.*, **69**, 514.

SAUER, R., 1947, *Conference on Non-stationary Gas Dynamics* (Laboratoire d'Etude Ballistiques de Saint Louis).

SMITH, L. G., 1945, *Declassified O.S.R.D. report* 6271.

SOMMERFELD, A., 1895, *Math. Ann.*, **47**, 317.

TAUB, A. H., 1947, *Phys. Rev.*, **72**, 51 ; 1948, *Phys. Rev.*, **74**, 328.

TAYLOR, G. I., 1941, *Ministry of Home Security reports* (unpublished); 1946, *Proc. Roy. Soc.* A, **186**, 273.

TAYLOR, G. I., and MACCOLL, J. W., 1935, *The Mechanics of Compressible Fluids*, Ed. W. F. Durand, Vol. VIII (Berlin : Springer).

THOMAS, L. H., 1944, *J. Chem. Phys.*, **12**, 449.

UNWIN, J. J., 1941, *Proc. Roy. Soc.* A, **178**, 153.

WEISE, A., 1943, various publications by Institüt für Gasdynamik der deutschen Versuchsanstalt für Luftfahrt, some of which have been translated by B.I.O.S. Group 2.

ZUCKERMAN, S., 1940, *The Lancet*, 219; 1941, *Proc. Roy. Soc. Med.*, **34**, 171, *Brit. Med. J.*, **1**, 645, *Ministry of Home Security report* (unpublished).

ZUCKERMAN, S., *et al.*, 1942, *The Lancet*, 252.

FERROMAGNETISM: MAGNETIZATION CURVES

By EDMUND C. STONER
Department of Physics, The University, Leeds

CONTENTS

	PAGE
§ 1. Introduction	83
§ 2. Single crystals	85
(i) Introductory	85
(ii) Experimental methods and results	86
(iii) Theory of magnetization curves	97
(iv) Atomic theory	103
§ 3. Domain structure	107
(i) Theoretical and general	107
(ii) Powder patterns	117
(iii) Discontinuities in magnetization: the Barkhausen effect	123
(iv) Related investigations	129
§ 4. The analysis of magnetization curves	132
(i) Introductory	132
(ii) Boundary movements	134
(iii) Reversible rotational processes	141
(iv) Rotational processes and hysteresis	145
(v) Magneto-thermal effects	150
§ 5. Magnetic materials	157
(i) Introductory	157
(ii) General survey	158
(iii) Iron–silicon and nickel–iron alloys	163
(iv) Permanent-magnet materials	169
§ 6. Conclusion	177
Acknowledgments	179
References	180

§ 1. INTRODUCTION

This report, dealing mainly with work since 1934 on the variation with field of the magnetization of ferromagnetics and associated phenomena, concludes a survey of ferromagnetism of which the first section (Stoner 1948) appears in an earlier volume of these *Reports*. To quote from the previous report (which will be referred to as I): 'The subject matter which might appropriately be considered under the title "Ferromagnetism" has been thought of as falling into three parts: the first, the general ideas and principles gained from the work up to about 1934, giving what may be called the "Background" to the more recent developments; the second, the theoretical and experimental work since that date connected with the first major theme, "Intrinsic Magnetization"; and the third, the more recent work connected with the second major theme, which can be indicated by the title "Magnetization Curves".' The first two parts were covered in I, in which the introduction and account of the background (I, pp. 43–65) are to be regarded as prefatory to the present account of magnetization curves as well as to the earlier account of intrinsic magnetization (I, pp. 65–108). As it is clearly impracticable in a report of this kind to deal adequately with work since 1934 on all aspects of the field covered by the title 'Magnetization Curves',

it is perhaps desirable first to indicate the general approach and the criteria adopted in the selection of subject matter and in deciding on the degree of detail with which it is to be reviewed.

It has long been established that a ferromagnetic below the Curie point is magnetized even in the absence of an applied field. The spontaneous magnetization is, however, not in general uniform throughout a specimen, but varies in direction from one region to another, the regions of unidirectional magnetization usually being referred to as 'domains'. At temperatures well below the Curie point (e.g. at room temperature for iron, cobalt and nickel), as the 'apparent' magnetization varies with applied field from zero to 'saturation', the intrinsic magnetization changes little, and the spontaneous magnetization in zero field is approximately equal to the saturation magnetization as ordinarily measured. Ferromagnetics are characterized essentially by the magnitude and temperature variation of the spontaneous magnetization up to the Curie point, and above the Curie point by the magnitude and temperature variation of the susceptibility. The fundamental theoretical problem of ferromagnetism is to account for the occurrence of spontaneous magnetization and for its magnitude in different elements, alloys and compounds. A reasonably full survey has already been given of both the theoretical (I, pp. 65–78) and experimental work (I, pp. 78–108) since 1934 primarily concerned with intrinsic as distinct from apparent magnetization. The occurrence of spontaneous magnetization having been satisfactorily explained in terms of relatively strong exchange interaction forces tending to align the elementary carriers of the magnetic moment (in metallic ferromagnetics, electron spins), the basic problem presented by the field variation of apparent magnetization is to account for the non-attainment of saturation in relatively weak, even vanishingly small, applied fields. It is worth noting in this connection that the field required to produce approximate saturation, even if it is as high as several thousand oersteds, is still very much less than the effective field (the molecular field) corresponding to exchange interaction, which is of the order of 10^7 oersteds. The whole complex of behaviour represented by the magnetization curves of ferromagnetics may thus be regarded as arising from small perturbations. The sorting out and quantitative characterization of the main perturbing effects is essential to the understanding of magnetization curves, and particular consideration will be given to experimental and theoretical work concerned with this aspect of the subject.

The greater part of experimental work on ferromagnetism, stimulated in no small degree by the technological requirement for materials with particular magnetic properties or combinations of these with other properties, is concerned with the measurement of the variation in low and moderate fields of the apparent magnetization of different materials, including innumerable commercial alloys and special laboratory samples prepared in the course of their development. There is astonishing variety in these magnetization curves, as exemplified by the fact that materials whose saturation magnetization is much the same may have coercivities ranging from a few hundredths of an oersted to several hundred oersteds, and maximum permeabilities ranging from less than a hundred to more than a million. It is not intended in this report to provide a compilation of results on the many materials which have been examined, but reference will be made to sources where such data are collected. The intention is rather to select for consideration those magnetic and associated investigations which provide

quantitative information about the elementary processes of magnetization, which exemplify general methods of attack on representative problems, or which illustrate general principles. It is kept in mind that a dominant aim in research on ferromagnetism is the determination, with such precision as may be possible, of just what is happening on a microscopic scale, sub-crystalline or atomic, when the state of bulk magnetization changes in any material, and the linking up of the magnetic properties with other properties and with what may be known or presumed about the microscopic constitution of the material.

As the various topics to be considered have close interconnections, a fair amount of cross-reference between the four main sections, as well as between the sub-sections, is unavoidable. To reduce forward reference to a minimum, the order of presentation is very roughly one of decreasing generality. The work on single crystals (§ 2) gives clear-cut information about one type of magnetization change, by the rotation of domain magnetization vectors, which is of general occurrence. A fully satisfactory quantitative interpretation has only recently been developed of the simplest types of magnetization curves, those of single crystals over effectively a purely rotational range, treating the magneto-crystalline anisotropy coefficients as parameters to be determined experimentally. This is an indication of the complexity of the problem presented by magnetization curves generally. A satisfactorily quantitative 'explanation', on an atomic theory basis, of the magnitudes of these anisotropy coefficients is still far from having been obtained. The investigations on domain structure (§ 3) theoretically, and experimentally by the powder pattern method, and through the study of the Barkhausen effect, again give information which is very wide in its bearings. The sections on single crystals and domain structure provide a basis for a more detailed consideration of the analysis of magnetization curves generally (§ 4). Accounts are given of some of the more recent work on reversible and irreversible domain boundary movement and rotational processes, on the contribution of these to change of magnetization along magnetization curves, and on their relation to such characteristics as permeability, coercivity and remanence. A short general survey of ferromagnetic materials is followed by slightly fuller accounts of work on a few representative materials, chosen in part to illustrate how various types of technological requirement are being met (§ 5). In some concluding remarks (§ 6), attention is drawn to what are believed to be the main lacunae in the report within the field which the title might be presumed to cover, and comment is made on the relation between this report and various books and general articles which have appeared during the period under review.

§ 2. SINGLE CRYSTALS

2 (i). *Introductory*

Work on the magnetic properties of single crystals up to 1934 (I, pp. 58–61 has been very fully reviewed by McKeehan (1934 a). This work included measurement of the magnetization curves with the field along the main axial directions for iron, cobalt and nickel at room temperature, and for iron and cobalt over a wide temperature range, and measurement of torque curves (giving essentially the component of magnetization normal to the field) for single crystal discs, and a fairly complete development of the formal theory of magneto-crystalline anisotropy. Since 1934 the experimental techniques have been improved, and the modified methods have been applied to single crystals of alloys as well as of the ferromagnetic

elements; treatments of the magnetization curves applicable to the lower as well as the higher magnetization ranges have been developed; and further attacks have been made on the atomic theory of magneto-crystalline anisotropy. The work will be reviewed under three corresponding headings.

Expressions consistent with the crystal symmetry for the part of the (internal) free energy dependent on the direction of the magnetization may first be recalled (cf. I, pp. 59, 60). For a cubic crystal the expression, including only terms which are practically significant, may put be in the form

$$F_c = K_4(\alpha_2^2\alpha_3^2 + \alpha_3^2\alpha_1^2 + \alpha_1^2\alpha_2^2) + K_6\alpha_1^2\alpha_2^2\alpha_3^2, \quad \ldots\ldots(2.1)$$

where the α_r are direction cosines of the magnetization, I_0, referred to a coordinate system defined by the cubic axes. For a hexagonal crystal which is isotropic in the basal plane (as is hexagonal cobalt, with close approximation) a convenient form of expression is

$$F_c = K_2' \sin^2\psi + K_4' \sin^4\psi, \quad \ldots\ldots(2.2)$$

where ψ is the angle between I_0 and the hexagonal axis. The anisotropy coefficients, as used here, are to be regarded as defined by the above equations, which are those most commonly used. It has seemed desirable to give these equations explicitly, since by adding or subtracting constants (including functions of the angles which are constant), various alternative expressions may be obtained, in which the coefficients will have different numerical values; and it not infrequently happens in the literature that numerical values are quoted for anisotropy coefficients which are incorrect for the form of the free energy expression which may be implied or actually given. It may be noted that in a cubic crystal, unless K_6 is numerically greater than $9K_4/4$, and of opposite sign, if K_4 is positive, as for iron, the energy is a minimum for magnetization along a cube edge direction; the cube edges are directions of easy magnetization, and in the demagnetized state the domain magnetization vectors will be equally distributed along the six equivalent cube edge directions. With K_4 negative, as for nickel, the easy directions are the cube diagonals. For cobalt, the easy direction is along the hexagonal axis, and K_2' is positive. Reference will be made below to the methods, largely developed before 1934, for determining the anisotropy coefficients from the experimental magnetization and torque curves.

2(ii). *Experimental Methods and Results*

The experimental measurements on single crystals are essentially of the components of magnetization parallel and normal to the field, I_p and I_n, as functions of the field for different orientations of the field relative to the crystal axes. In principle, leaving aside the low field region, full information for cubic crystals (i.e. the values of K_4 and K_6 in (2.1)) can be obtained from I_p curves for the field along the three directions [100], [110] and [111], for which I_n is zero. Such measurements have been made by ballistic methods, particularly by the Japanese investigators, the search coil arrangements being adapted to the small monocrystalline rods or discs of suitable crystal orientation which are available. (A typical size of 'rod' is 2 cm. long by 0·01 cm^3 cross section.) Ballistic methods have also been used for determining the normal component of magnetization, I_n, but this is usually obtained from measurements of the couple on a circular disc, or preferably an oblate spheroid, suspended horizontally in a horizontal field, with the equatorial plane as near as possible to a simple crystal plane. From

curves (torque curves) giving the couple against field orientation, the anisotropy coefficients can again in principle be determined most readily if results are available for a series of high field values. (As the field increases, the torque tends asymptotically to a limit, and the calculations are greatly simplified if these limiting torque curves can be estimated with precision.) Torque methods have also been used to determine I_p by suspending the disc vertically, and measuring the couple required to maintain it at a small angle to the field, but the corrections are difficult to estimate with adequate precision.

Torque and ballistic methods. A number of torque magnetometers have been developed since 1934 for the study both of single crystals and of rolled sheet material. A simple and sensitive form has been fully described by Williams (1937 a), a more robust balance arrangement with a rotating magnet by Tarasov and Bitter (1937) (see also Tarasov 1939), and an instrument specially suited to sheet material by Brailsford (1938). The experimental determination of torque curves is fairly straightforward, and there is no fundamental difficulty in the analysis of the results. When the ballistic method is used with rod-shaped specimens there are unavoidable uncertainties in connection with demagnetizing effects. These uncertainties have been ingeniously avoided by Williams (1937 b), who has succeeded in cutting single crystals in 'picture frame' form with the sides along tetragonal or digonal directions in a (100) plane, and along trigonal directions in a (110) plane. The picture frames are small (the sides of a rectangular tetragonal (100) specimen had mean lengths of 1·72 and 1·23 cm., and the cross-sectional area was about 0·07 cm^2), but they are large enough to take adequate primary and secondary windings as in the classical anchor ring arrangement, and reliable and accurate measurements can be made in low fields (below 0·1 oersted) as well as in high.

The pendulum magnetometer. A new method for the study of single crystals has been introduced by McKeehan by a development of the Weiss–Foëx translational balance, previously used only for dia- and paramagnetics (McKeehan 1934 b, McKeehan, Piety and Kleis 1936). The 'pendulum magnetometer', as the instrument is called, consists of a light bar supported by a quinquefilar suspension in such a way that it is constrained to move only in the direction of its length and without rotation. The bar, on which the small ferromagnetic specimen is suitably mounted, passes through the solenoid producing the main field so that the specimen is on the axis. The solenoid may be moved along its length, and is so placed that the specimen is central. Two coaxial coils (a Helmholtz type of arrangement), separated by spacers, can slide as a unit over the solenoid. Traversed by the same current in opposite directions. these coils provide a field gradient. If I_p is the component of magnetization of the specimen along the direction, say x, of the bar, the force per unit volume is given by

$$f_x = I_p(\partial H/\partial x),$$

and $(\partial H/\partial x)$ can be adjusted independently of the main field H. Corrections are made for lack of uniformity of the field over the small range of displacement of the specimen. The force is measured by the deflection, suitably magnified, of the pendulum from its equilibrium position in the earth's gravitational field, observation being made of the movement on a distant scale of the image of an illuminated cross wire, or, as described by Grabbe (1940), a portion of a replica grating, mounted on the pendulum bar. Overall sensitivities of the order

100 mm. dyne^{-1} have been obtained. The specimens are prepared in the form of small oblate spheroids. (It is to be noted that for homogeneous material in a uniform field, the magnetization is also uniform for an ellipsoid, and the demagnetizing factors may be expressed, in the general case, in terms of elliptic integrals. For ellipsoids of revolution, i.e. oblate or prolate spheroids, the demagnetizing factors are expressible in terms of elementary functions. The evaluation is straightforward, but tiresome. Tables of values which would be useful in the present connection, and others, are given by Stoner (1945). It is surprising how often calculated values in the literature are in error to a physically significant extent.) In preparing the spheroid, a disc-shaped blank is mounted on the end of a vertical shaft of a variable slow speed motor, and cut by fine files while rotating. A magnified image is thrown on a screen, and the cutting operation continued until the silhouette fits accurately a carefully plotted ellipse of the required form. Typical spheroids have a thickness (polar diameter) of about 0·2 mm., and an equatorial diameter of about 4 mm. Etch reflection and x-ray methods are used for determining the crystal orientation of the initial single crystal blanks and of the final spheroids. The most convenient orientation for the equatorial plane is (110), as this includes the three main axial directions. Appropriate annealing treatments are applied to remove strains. The pendulum magnetometer has been successfully applied to the determination of the single crystal magnetization curves (I_p, H with the field along different axes) of iron and of a large number of alloys by McKeehan and his co-workers Shih, Piety, Kleis and Grabbe, referred to below.

Derivation of anisotropy coefficients. In the treatment of single crystal magnetization it has usually been assumed that in the demagnetized state the domain magnetization vectors are equally distributed along the equivalent easy directions of magnetization, and that, as the applied field in a particular direction is increased, domain boundary movements take place until, still in a relatively small field, all the domain magnetization vectors are pointing along the easy direction nearest to the field. There is then a single magnetization vector to be considered, and this, as the field increases further, rotates gradually towards the field direction. The earlier formal theory was mainly concerned with this 'single-vector treatment' as it has been called (von Engel and Wills 1947), which is usually taken to include the special cases in which the field direction is symmetrically between two or three easy directions (e.g. the field along [110] or [111] for cubic crystals with K_4 positive), as then it is necessary to consider only one of the two or three magnetization vectors which are equal in magnitude and make equal angles with the field. The deficiencies, both recognized and unrecognized (see § 2 (iii) below), of this treatment do not appreciably affect the validity of the methods which have usually been applied to the derivation of the anisotropy coefficients.

The first and probably most widely used method is from energy density differences, obtained from the measurement of areas between the I axis and the magnetization curves for different directions. The increase in internal energy in change of magnetization from zero to saturation may be written

$$W = \int_0^{I_s} H dI. \qquad \ldots\ldots (2.3)$$

For a cubic crystal the value of W for magnetization in any given direction is obtainable from the difference between the values of F_c in (2.1) for the final and

initial values of the α_r. To allow for the effect of random irregularities, assumed to affect the magnetization curves in all directions equally, differences are taken, giving

$$\left.\begin{aligned} W_{110}-W_{100}&=K_4/4,\\ W_{111}-W_{100}&=K_4/3-K_6/27,\end{aligned}\right\} \quad \ldots\ldots(2.4)$$

or

$$\left.\begin{aligned} K_4&=4(W_{110}-W_{100}),\\ K_6&=9\{4(W_{110}-W_{100})-3(W_{111}-W_{100})\}.\end{aligned}\right\} \quad \ldots\ldots(2.5)$$

Some magnetization curves for iron, copied from those given by Piety (1936) as obtained by the pendulum magnetometer method, are shown in Figure 2.1. In so far as these curves are a 'fair' representation of the experimental results, it would seem that reliable values for the required areas could be obtained from the data in this particular work.

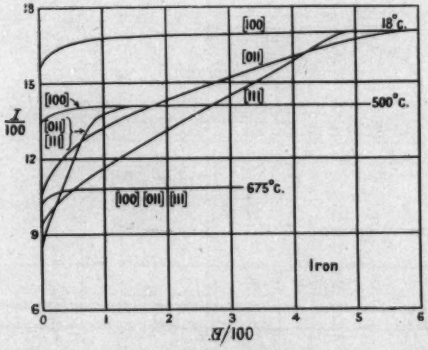


Figure 2.1. Magnetization curves for single crystals of iron at 18°, 500° and 675° c. *After* Piety (1936).

In Figure 2.2 some representative curves for nickel are shown. These have been drawn smoothly through points plotted from the numerical data of Honda, Masumoto and Shirakawa (1935), obtained by the ballistic method, using monocrystalline rods.

It is clear, to some extent from the figure itself, that the values of the anisotropy coefficients obtained from an evaluation of the areas, which has been carried out by McKeehan (1937 a, p. 139), cannot here be particularly reliable. It is very questionable whether departures from the ideal strain-free state can be allowed for satisfactorily by taking differences, as in (2.4), when different specimens must be used in determining the magnetization curves in different directions. Further, unless particular care has been taken in connection with the measurements in the higher fields, there may be large errors in the estimates of the areas up to saturation.

The form of the magnetization curve for a particular direction of the field may be calculated for the single-vector region by minimizing the total free energy (obtained by adding to F_c the external field energy, $-HI_0\cos\phi$, where ϕ is the angle between H and I_0) with respect to the direction of I_0. By appropriate choice of the anisotropy coefficients good general agreement with the observed curves can usually be obtained. In view of experimental irregularities and the imperfections of the specimens, however, the criterion of 'best fit' cannot usually be satisfactorily applied for a primary determination of the coefficients, for which a more clear-cut method is desirable.

The basic theory of the torque curve method of investigating single crystals, due mainly to Akulov (1931) and Gans (1932), has been developed in considerable detail by Bozorth (1936), Schlechtweg (1936), Tarasov and Bitter (1937) and Bozorth and Williams (1941). The general theory of the torque on a spheroid

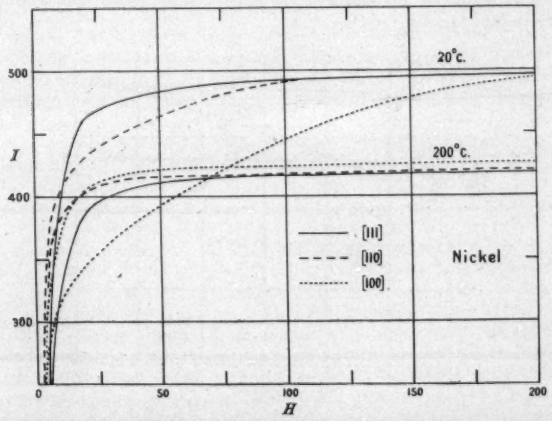


Figure 2.2. Magnetization curves for single crystals of nickel at 20° and 200° c. *From the data of* Honda, Masumoto and Shirakawa (1935).

when the equatorial plane has any crystal orientation is very complicated geometrically (or trigonometrically) even when the single vector assumption is made. If the aim is primarily the determination of the anisotropy coefficients, attention may be restricted to the equatorial orientations (100), for which the torque curves depend on K_4 only, and (110), for which the curves depend on both K_4 and K_6. These cases, with others, are dealt with by Bozorth (1936), and it will be sufficient to take one particular case in illustration. For a (110) spheroid, I_0 is confined (except in small fields) to the equatorial plane. Denoting the angles between a chosen reference direction, say [001], and H and I_0 by θ and ψ respectively, the torque, T, is given by

$$T=-\partial F/\partial\theta, \qquad \text{.....(2.6)}$$

where

$$F=F_c-HI_0\cos(\theta-\psi), \qquad \text{.....(2.7)}$$

and F_c is given by (2.1). Further, the equilibrium direction of I_0 is given by

$$\partial F/\partial\psi = 0. \qquad \ldots\ldots (2.8)$$

From these equations it may be shown without difficulty that

$$T = -(K_4/8)(2\sin 2\psi + 3\sin 4\psi) - (K_6/64)(\sin 2\psi + 4\sin 4\psi - 3\sin 6\psi). \qquad \ldots\ldots (2.9)$$

Now in this equation the torque is expressed in terms of the angle, ψ, not directly known, between I_0 and the reference direction, rather than in terms of θ, the known angle between H and the reference direction. From (2.8) and (2.7) a manageable expression is obtainable for θ as a function of ψ, but not in general for ψ in terms of θ. The complications are evaded (cf. Bozorth 1936) by the consideration that as H increases indefinitely, the direction of I_0 approaches that of H, and $\theta - \psi$ tends to zero. In (2.9), in the limit, ψ may be replaced by θ. Good agreement can be obtained between the experimental results in high fields, and the torque curve calculated from (2.9) in this way, adjusting the coefficients to give the closest fit, as illustrated in Figure 2.3, showing the results of Williams (1937 b) for a disc (1·77 cm. in diameter, 0·222 cm. thick) of 3·85% silicon iron.

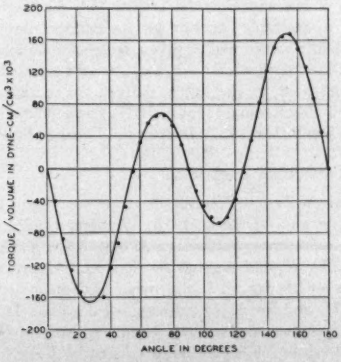


Figure 2.3. Torque curve for a single crystal disc of silicon iron in the (110) plane, taken with an applied field of 5,800 oersteds.

The line is drawn according to equation (2.9), with $K_4 = 28{\cdot}7 \times 10^4$, $K_6 = 10{\cdot}0 \times 10^4$ (Williams 1937 b).

(*Reproduced, by kind permission, from* ' *The Physical Review* '.)

It has been found by Tarasov (1939) that the maximum torque in a field H, which may be denoted by T_H, can be closely represented over the higher field range by

$$T_H = T_\infty(1 - H_0/H_a), \qquad \ldots\ldots (2.10)$$

where H_a is the applied field, and H_0 an empirical constant. Later work by Bozorth and Williams (1941) shows that this equation is not generally valid, but that, applied with discretion to high field results, it will usually give a good approximation to the limiting torque.

The above outline gives the essentials of the main methods which have been used for the determination of anisotropy coefficients, and it is unnecessary to consider the minor variants. Experimental investigations to test whether the different methods yield consistent results are clearly desirable. Such an investigation has been made by Williams (1937b), who determined the coefficients for 3·85% silicon iron by three methods: from the (I_p, H) curves obtained with picture frame crystals; by measurement of areas and by curve fitting; and from torque measurements (Figure 2.3). The values obtained for K_4 (in units of 10^5 erg cm^{-3}) were 2·72, 2·80 and 2·87; and for K_6, 1·5, 1·0 and 1·0. The agreement must be regarded as extremely satisfactory and would seem to indicate that different methods do give consistent results for a particular material when the experimental work and the analysis of the results are carried out with sufficient care. In general, however, there is a wide spread in the values given for the anisotropy coefficients as derived from the data (either by the experimentalists themselves or by others) on materials differing little in composition, and similarly heat treated. It will be clear from (2·5) that comparatively small experimental errors may result in large errors in the estimated values of K_6, and this is probably sufficient to account for the very different values given. For K_4 the agreement is closer, but the range of values is still much wider than would be expected from the estimated accuracy of the experimental results. A thorough critical survey of the data, such as is desirable, cannot be undertaken here. Rather crudely estimated most probable values will be given, with some indication of the spread. General surveys of the experimental work, and lists, of varying degrees of completeness to the dates indicated, of the values derived for the anisotropy coefficients are given in papers by McKeehan (1934a, 1937a) and Bozorth (1937) and in the books by Stoner (1934), Bitter (1937), Becker and Döring (1939) and Bates (1950).

Numerical values of anisotropy coefficients.

Iron. McKeehan (1934a, p. 27) lists the values he obtains for $W_{110} - W_{100}$ and $W_{111} - W_{100}$ from the data of Beck (1918), Webster (1925), Honda and Kaya (1926) and Honda, Masumoto and Kaya (1928). From these energy density differences, using (2.5), K_4 and K_6 may be derived. Expressing the anisotropy coefficients in units of 10^5 erg cm^{-3}, the mean and mean deviation are for K_4 (7 values) $4{\cdot}16 \pm 0{\cdot}53$, and for K_6 (5 values), $-1{\cdot}2 \pm 1{\cdot}9$. It may be noted that some of the same experimental data have been analysed by Gans and Czerlinski (1932, 1933), whose values for K_4 and K_6 are listed by Becker and Döring (1939, p. 123). From the same sets of experimental results for which McKeehan's analysis gives, as a mean of three values, 4·69 for K_4, Gans and Czerlinski obtain 4·35; and the means of two values of K_6 are, respectively, $-2{\cdot}7$ and $+2{\cdot}6$.

The later work on iron includes that of Piety (1936) by the pendulum magnetometer method (see Figure 2.1) and of Tarasov (1939), by the torque magnetometer method. For one specimen Piety obtained $K_4 = 4{\cdot}42$, $K_6 = +1{\cdot}4$ and for another $K_4 = 4{\cdot}27$, $K_6 = -1{\cdot}7$. It is obvious from these results, and those above, that no significant 'probable value' can be given for K_6 for iron. The K_4 values agree reasonably well with each other, and with the mean of the earlier values. Tarasov's torque measurements, however, on one of the spheroids which had been used by Piety, gave a value of 5·29, considerably higher than any previously listed, and for a second specimen 4·74. In an excellent discussion of his own and relevant previous work on iron, Tarasov concludes that the earlier torque

curve data, properly treated, are not incompatible with values of K_4 higher than those usually given, and comparable with his own; but he cannot account for the lower values obtained by the area method from I_p curves. The position is clearly very unsatisfactory, but until further investigations are made, the best that can be done is to take as a working probable value of K_4 the mean of the just quoted values of Piety and Tarasov, since the iron used was of high purity (better than 99·9%) and each investigator was aware of such possible sources of error as have so far been recognized in the method he used. This gives $K_4 = 4{\cdot}68 \pm 0{\cdot}34$, or, nearly enough, $K_4 = 4{\cdot}7$, with an uncertainty of at least 10%.

The most extensive investigation of the temperature variation of the (I_p, H) curves for different crystal directions is that of Honda, Masumoto and Kaya (1928), whose measurements covered a range from 5° C. to 770° C. Piety (1936)

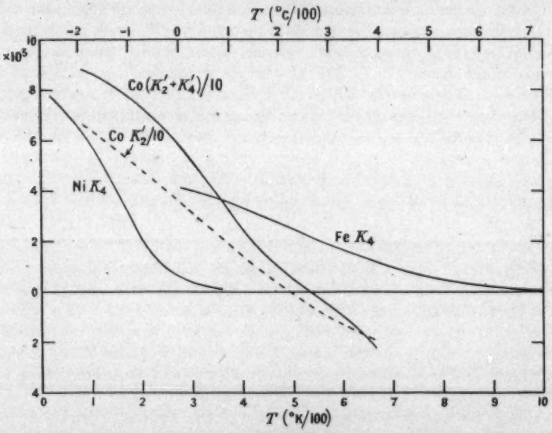


Figure 2.4. Temperature variation of the magneto-crystalline anisotropy coefficients of iron, nickel and cobalt.

Based on the experimental data of Honda, Masumoto and Kaya (1928) for iron, of Brukhatov and Kirensky (1937) and of Williams and Bozorth (1939) for nickel, and of Honda and Masumoto (1931) for cobalt. For definitions of K_4, K_2' and K_4' see equations (2.1) and (2.2). The coefficients are expressed in erg cm^{-3}. The methods of evaluation are described in the text.

has evaluated K_4 (and also K_6) from their results, and gives a list of values obtained by interpolation for conveniently spaced temperatures over the range. A curve drawn through points plotted from his values is shown in Figure 2.4. Piety's own results give systematically higher values for K_4 at 18°, 296° and 500° C., but are consistent with a curve of the same form. At 675° C. (see Figure 2·1) and 696° C., to the accuracy of the measurements, his magnetization curves for the [100], [110] and [111] directions coincide, corresponding, formally, to a zero value of K_4; but a tailing off, as in the curve shown, is not precluded. A curve given by Titow (1936), showing the relative values of K_4 over the range 0° to 450° C., as obtained from torque measurements, is of the same general form as that in Figure 2.4.

Nickel. The earliest measurements on single crystals of nickel are those of Sucksmith, Potter and Broadway (1928) and of Kaya (1928). The values of the anisotropy coefficients which have been obtained from them are remarkably consistent. The areas obtained by McKeehan (1934 a) from the (I_p, H) curves correspond in both cases to $K_4 = -0{\cdot}51$ (in units of 10^5 erg cm^{-3}). Gans and Czerlinski (1933) derive the same value from Kaya's (I_p, H) curves, by curve fitting, and their values from four sets of (I_n, H) curves of Sucksmith *et al.* correspond to $K_4 = -0{\cdot}48 \pm 0{\cdot}05$. The various results are fitted reasonably well without taking the K_6 term into account, being consistent with a value of K_6 numerically less than 0·1.

The later work of Honda, Masumoto and Shirakawa (1935), covering a range of temperature from $-252°$ to $+370°$ C., has already been mentioned (see Figure 2.2). From an evaluation of areas McKeehan (1937 a) obtains the very different room temperature values $K_4 = -0{\cdot}34$, $K_6 = +0{\cdot}53$. The results of Brukhatov and Kirenski (1937), obtained by the torque method, using a spherical specimen, cover the range from 77° to 390° K. At room temperature (taken here, for definiteness, as 290° K.) their value of K_4 is $-0{\cdot}50$. (No values are given for K_6.) They find that, apart from a slight deviation over the upper part of the range covered, the results are represented closely by

$$K_4 = (K_4)_0 \exp(-aT^2),$$

where $$(K_4)_0 = -8{\cdot}0 \times 10^5, \qquad a = 3{\cdot}4 \times 10^{-5}.$$

At the lower temperatures the magnetization curves of Honda *et al.* indicate enormously higher values than those given by the above equation. Torque measurements by Williams and Bozorth (1939) gave 1·2 (to an accuracy of within about 10%) as the ratio of the values of K_4 at 20° K. and 77° K. This agrees with the ratio of 1·21 given by the equation, while the Honda results correspond to a ratio of about 5. At 77° K. and above Williams and Bozorth obtained absolute values which were very close to those of Brukhatov and Kirenski, whose results are shown graphically in Figure 2.4. The change in order of ease of magnetization along different axes indicated by the Honda curves (Figure 2.2) is very striking, but it is clearly unjustifiable to attach much weight to the numerical estimates of anisotropy coefficients which may be obtained from them.

The salient facts to be noticed are that at room temperature K_4 for nickel is numerically about one tenth of K_4 for iron and of opposite sign; and that the temperature variation is more pronounced, but of somewhat similar character. For neither element can a satisfactory estimate be made at present of the value of K_6 for an 'ideal' crystal.

Cobalt. The only measurements so far available on single crystals of cobalt are those of Kaya (1928), at room temperature, and of Honda and Masumoto (1931), over a temperature range from $-190°$ to $+390°$ C. (The stable form of cobalt is hexagonal below and face-centred cubic above a temperature of about 470° C.) At room temperature the easy direction of magnetization is that of the hexagonal axis, [0001], for which the magnetization rises almost to the saturation value in low fields. For directions in the basal plane, (0001), the magnetization increases slowly, and the curve, which is concave towards the H axis, can be represented reasonably well by an equation derivable from (2.2) with a suitable choice of

K_2' and K_4'. The curves are slightly different along the axes of closest packing of the rows of atoms, typified by $[11\bar{2}0]$, and the axes bisecting these, $[10\bar{1}0]$, the curves for the latter being slightly higher (magnetization is 'easier') than for the former; but the energy difference is very small compared with that between either and the hexagonal axis, and will not be considered here. If isotropy in the basal plane is assumed, it will be clear from (2.2) that the sum $K_2' + K_4'$ gives directly the energy density difference for magnetization along and normal to the hexagonal axis. This may be directly derived by the area method from (I_p, H) curves for the two directions. The determination of K_2' and K_4' separately is possible only by a curve fitting method.

From Kaya's results the area evaluations of McKeehan give at room temperature $K_2' + K_4' = 52$ (in units of 10^5 erg cm^{-3}), while Gans and Czerlinski obtain for the separate coefficients $K_2' = 41$, $K_4' = 10$. The values of $K_2' + K_4'$ corresponding to the areas evaluated by McKeehan from the curves of Honda and Masumoto are plotted in Figure 2.4. The value at room temperature is about 59. For K_2' and K_4' separately the literature provides a wide choice (see, for example, Bozorth 1937, Van Vleck 1937, p. 1198). The curve for K_2' in Figure 2.4, which is plotted from values given by Bitter (1937, p. 212), who has made 'slight alterations and corrections' to values given earlier by Gans and Czerlinski, must therefore be regarded as having a lower degree of certainty than that for the sum of the two coefficients. The reliability of the sum curve must be assessed from the fact that values of 52 and 59 are obtained at room temperature from two different sets of results.

The magneto-crystalline anisotropy coefficients of the ferromagnetic elements have been considered in detail (a detail much greater than will be possible for other topics in this report) mainly because a knowledge of the values of these coefficients is a basic requirement in any quantitative treatment of magnetization curves of polycrystalline as well as monocrystalline material, and the uncertainty of the values is not generally appreciated. There is, however, no doubt about general orders of magnitude, and it may be well to summarize the broad facts, which is conveniently done by reference to the difference between the energy densities for magnetization in the easiest and most difficult directions. For cubic crystals this difference is given with sufficient accuracy by $K_4/3$, and for hexagonal crystals by $K_2' + K_4'$. The values at room temperature for nickel, iron and cobalt are about $1{\cdot}7 \times 10^4$, $1{\cdot}4 \times 10^5$ and $5{\cdot}9 \times 10^6$ respectively. Much of the diversity of the 'ordinary' magnetic properties of the three metals can be linked with the differences in these basic characteristics. The maximum energy density difference for nickel increases considerably below room temperature and the trend of the curves of Figure 2.4 suggests that it would be greater than that for iron, but of the same order of magnitude, at the lowest temperatures, the value for cobalt maintaining its relatively high magnitude. The values of the anisotropy coefficients at room temperature only give a quite misleading impression of the problem confronting the atomic theory of these effects.

Alloys. Measurements on single crystals of binary alloys of the ferromagnetic elements have been made by the pendulum magnetometer method by Shih (1934, 1936) for iron–cobalt, and nickel–cobalt, and by Kleis (1936) for iron–nickel. Eight ternary alloys in the face-centred region have been examined by McKeehan (1937a), who summarizes the main results on both the ternary and

binary alloys in two figures, reproduced in Figure 2.5, showing the derived values of K_4 at 20° c. and 200° c.

The main features of the results for K_4 are brought out so clearly in McKeehan's diagrams that no verbal description is necessary, but it should be noted that the K_4 values for nickel differ from those plotted in Figure 2.4, as they are derived from the earlier work of Honda, Masumoto and Shirakawa. The estimated values of

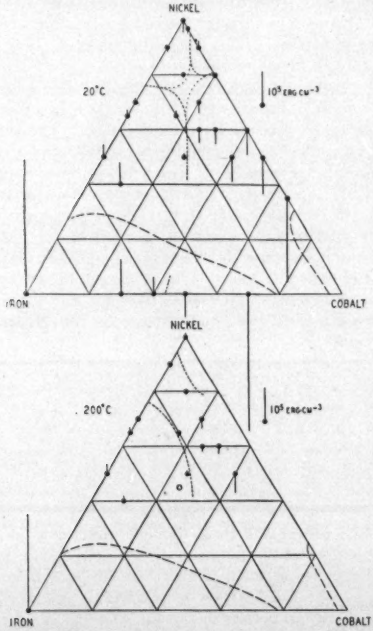


Figure 2.5. Principal anisotropy coefficient, K_4, of nickel–iron–cobalt alloys at 20° c. (upper figure) and 200° c. (lower figure), as determined from the results of Shih (Fe–Co and Ni–Co), Kleis (Ni–Fe) and McKeehan (ternary alloys).

Lines up or down from a dot indicate positive or negative K_4, according to the scale on the right. Broken lines show approximate phase boundaries (hexagonal round cobalt, body-centred cubic round iron, face-centred cubic otherwise). The dotted curves are provisional estimates of the zero K_4 lines. (McKeehan 1937 a.)

(*Reproduced, by kind permission, from* ' *The Physical Review* '.)

K_6, which, as is stated, are very uncertain, show no obvious regularities. None the less, as has been stressed particularly by Bozorth (1936, 1937), a satisfactory formal coordination of the observed magnetization curves cannot in general be obtained unless account is taken of the K_6 term; this is shown most clearly by the occurrence of orders of ease of magnetization other than [100], [110], [111] or the converse in certain composition or temperature ranges.

The results of a detailed investigation of the anisotropy of iron–nickel alloys with from 65 to 80% nickel by Grabbe (1940), using the same method, are most appropriately left for consideration in a later section (§ 5 (iii)).

Extensive measurements have been made by the torque magnetometer method on iron–silicon alloys with an atomic percentage, A, of silicon up to 13·7 by Tarasov (1939), who finds that the values he obtains for K_4 (in units of 10^5 erg cm^{-3}) are closely represented by

$$K_4 = 5{\cdot}29 - 0{\cdot}279\,A \quad (A < 9{\cdot}86),$$

$$K_4 = 4{\cdot}43 - 0{\cdot}192\,A \quad (A > 9{\cdot}86).$$

A later value for K_4 obtained by Bozorth and Williams (1941) for a 3·6% specimen (atomic per cent, A, 6·9) is 3·5, about 7% higher than the value given by Tarasov's first equation. Williams's earlier values, mentioned above, for a 3·85% specimen, are about 14% lower.

It may be noted that although added silicon reduces the anisotropy coefficient of iron considerably (1% by weight gives a reduction of about 10%) the effect is not so large as to suggest that more than a small part of the spread of the 'observed' values for ordinarily pure iron can be accounted for in terms of the presence of impurities.

Miscellaneous. A considerable number of investigations have been made on diverse magnetic characteristics of single crystals, the results of which are peculiar to particular materials and specimens rather than being of general relevance. Bare reference must therefore suffice to work on the variation with direction of initial permeability by Williams (1937 c), which has been discussed by Bozorth (1937), Kondorski (1938), and Becker and Döring (1939, p. 133), and supplemented by futher experiments of Shoenberg and Wilson (1946); on the coercivity of iron–silicon by Sixtus (1937 b) and Jaanus and Schur (1937), and on the coercivity and maximum permeability of very pure iron and iron nickel (66% Ni) by Cioffi, Williams and Bozorth (1937); and on the hysteresis losses in iron–silicon (2.1% Si) by Wilson (1946).

2 (iii). *Theory of Magnetization Curves*

It will be clear from the discussion of the derivation of anisotropy coefficients from experimental data on single crystals that a simple formal treatment covers satisfactorily first the variation with field of I_p when H is along an axis of symmetry (I_n then being zero), and secondly the variation of I_n with both direction and magnitude of the field provided that the field is large. In view of the extensive application made of them, it may be useful to give the expressions, first derived by Akulov (1929, 1931) and Gans and Czerlinski (1932), for the magnetization with the field along the axes of symmetry. Writing $I_p/I_0 = j$, the expressions are as follows:

H along [100], $$j = 1, \qquad \ldots\ldots (2.11)$$

H along [110], $$j^3 - \tfrac{1}{2}j = HI_0/4K_4, \qquad \ldots\ldots (2.12)$$

H along [111], $$\frac{K_4}{3}\{j(7j^2-3) + 2^{\frac{1}{2}}(1-j^2)^{\frac{1}{2}}(4j^2-1)\} + \frac{K_6}{18}\{j(23j^4 - 16j^2+1) - 2^{\frac{1}{2}}(1-j^2)^{\frac{1}{2}}(10j^4 - 9j^2+1)\} = HI_0. \qquad \ldots\ldots (2.13)$$

Although it has been recognized for some considerable time that the variation of I_p and I_n in lower fields of arbitrary direction could not readily be reconciled with the theoretical scheme, it is only recently that a quantitative assessment of the apparent discordance has been made by von Engel and Wills (1947), and that a new method of approach to the problem, which removes most of the difficulties, has been developed independently by Néel (1944 a) and by Lawton and Stewart (1948). (It may be noted in passing that even the existence of papers published during the 1939–45 period could often not be known to many of those interested until much later.)

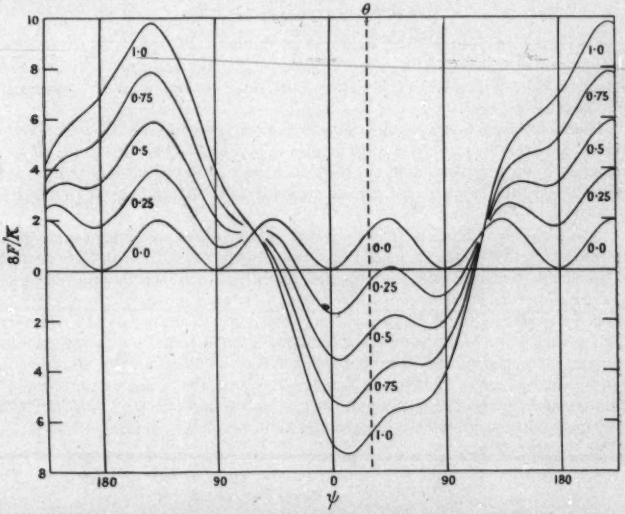


Figure 2.6. Relation between energy and direction of magnetization for a single iron-like crystal in a (100) plane. ψ, angle between [100] and I_0. $h=HI_0/K$. Angle, θ, between [100] and H, 30°. The numbers on the curves give the values of h.

Following the general treatment of von Engel and Wills, the essentials of the problem can be made clear by considering the dependence on direction of magnetization of the relevant part of the free energy of a uniformly magnetized iron-like crystal ($K=K_4$ positive, K_6 negligible; see (2.1)), with field and magnetization in the (100) plane. It is readily found from (2.1) that

$$F/K=\tfrac{1}{8}(1-\cos 4\psi)-h\cos(\theta-\psi), \qquad \ldots\ldots(2.14)$$

where ψ and θ are the angles between a [100] reference axis and I_0 and H respectively, and $h=HI_0/K$. Energy contours for various values of h are shown in Figure 2.6.

Of the four equal energy minima in zero field the one nearest to the field direction becomes more and more pronounced as the field increases, and moves towards the field direction. On the assumption that the domain magnetization vectors

come into alignment along the direction corresponding to the deepest energy minimum by domain boundary movements, it would be expected that a single vector treatment would cover the behaviour (that is, the variation of I_p and of I_n with H) except possibly in very low fields. Detailed calculations have been made by von Engel and Wills on this basis of the variation of the direction of I_0 (i.e. the value of ψ_{min}) with h for various values of θ, and of I_n/I_0 with θ for various values of h for the (100), (110) and (111) planes. The curves for I_n/I_0 against θ are, in general, surprisingly complicated, and reference must be made to the original paper for details. Roughly, it is only for values of h well above unity (corresponding, for iron, to values of H well above 280 oersteds) that the curves have a continuous periodic form similar to that found experimentally. The inadequacy of the theoretical treatment is shown most convincingly by a comparison of the theoretical curves for the variation of I_n/I_0 with h for particular values of θ with the curves derivable from experimental results (Webster 1925, Honda and Kaya 1926). For $\theta = 30°$, for example, as is apparent from Figure 2.6, I_n/I_0 should increase monotonically as h decreases, reaching the value $\sin\theta$ for zero h (or, strictly, for h just greater than zero). The experimental results are in agreement for high values of h (say $h > 10$), but, as h decreases, the experimental curve, diverging more and more from the theoretical curve, passes through a maximum, and approaches zero for zero field. It is to be noted that the direct experimental data are in terms of the applied field, H_e (e.g. a set of values may be available for I_n and I_p for H_e constant in magnitude and varying in direction), and the reduction of the data to a form appropriate to the use of H, the field 'corrected' for demagnetization effects, as the field variable, is, in general, both troublesome and uncertain. None the less, there is no doubt that the discrepancies between the single vector treatment and experiment are not in the main attributable to uncertainties arising in this way.

The experimental results in the lower fields are in fact inexplicable if the domain magnetization vectors are sharply restricted to the direction of a single deepest energy minimum. It was therefore proposed by von Engel and Wills that the domain directions might be distributed according to an expression of the form

$$dv_\psi = C \exp(-F/F_0)\, d\omega, \qquad \ldots\ldots (2.15)$$

where $dv_\psi/d\omega$ is the volume of domains per unit solid angle with I_0 at an angle ψ to the reference direction, F the corresponding energy per unit volume (given by an expression of the type (2.1)), and F_0 a constant reference energy. The single vector treatment already considered would correspond to the limiting case $F_0/K \to 0$. The larger the value of F_0/K, the wider the angular spread of domain magnetization directions. In low fields there might be an appreciable 'occupation' of directions not only near the lowest energy minimum but also near other low minima, but in high fields, as is required, only the lowest energy minimum need be considered. The particular experimental results examined by von Engel and Wills (those of Honda and Kaya on spheroids of iron) were fairly well (but by no means perfectly) covered, after correction for demagnetization, by taking for F_0 a value of about $K/8$. Now if the directional distribution (2.15) were of thermal origin, F_0 would be of the order kT/v_d where v_d is a suitably averaged single domain volume. With F_0 equal to $K/8$, this would give for the number of atoms per domain in iron about $200T$, i.e. about 6×10^4 at room temperature, a number which is too small by a factor of at least 10^8. Alternatively expressed, with

domains of the size indicated by other evidence, the 'thermal' spread of the directions of magnetization would be completely negligible in its effects. This was fully appreciated by von Engel and Wills, but they were unable to make any plausible suggestion as to the physical origin of a distribution parameter of the required order of magnitude. Unknown to them, a very general and detailed statistical treatment of the consequences of a formal assumption of the kind expressed by (2.15) had previously been developed by Takagi (1939), who also could find no physical justification of the magnitude apparently required for F_0. Since in the work now to be described very close agreement with experiment is obtained without any *ad hoc* assumptions in the theory, Takagi's extensive papers must be passed over here, though it should be stated that there is much of interest in the mathematical treatment, and that the various functions evaluated and tabulated may well have other fields of application.

The theoretical schemes of Néel (1944a) and of Lawton and Stewart (1948) for the interpretation of the magnetization curves of single crystals are in essentials identical, and it would be out of proportion to dwell here on interesting and characteristic differences in detail in the manner of approach. The simplest assumption of the domain theory is adopted, namely that the domain magnetization vectors lie along the direction of the lowest energy minimum, as determined

Table 2.1

Modes of Magnetization of Iron-like Single Crystals

The direction cosines of H, the resultant field, β_1, β_2, β_3, and of I_0, α_1, α_2, α_3, are with respect to the crystal axes.

Mode	Resultant field, H		Phases of magnetization	
	Magnitude	Direction cosines	No.	Direction cosines
I	$H=0$	—	6	$\langle 100 \rangle$
II	$H>0$	$\beta_1=\beta_2=\beta_3$	3	$[p, p, q]$, $[p, q, p]$, $[q, p, p]$
III	$H>0$	$\beta_1=\beta_2>\beta_3$	2	$[p, q, r]$, $[q, p, r]$
IV	$H>0$	$\beta_1>\beta_2>\beta_3$	1	$[p, q, r]$

jointly by the crystal anisotropy and the effective field, H, a resultant of the applied field, H_e and the demagnetizing field. There can therefore be two (or more) groups of domains with different directions of magnetization (constituting distinct 'phases', in Néel's terminology) only if there are two (or more) equal deepest energy minima. (In Figure 2.6, there are four such minima in zero field, but only one if the resultant field is not zero. If the angle, θ, between the [100] direction and the field were 45°, there would be two equal deepest minima.) Moreover, for the co-existence of two (or more) phases the resultant field direction must be symmetrical with respect to two (or more) easy axes. For an iron-like crystal (with cube edges as easy axes) there are four possible 'modes' of magnetization, corresponding to different numbers of distinct phases, as shown in Table 2.1 (cf. Néel 1944a, Lawton 1949).

The range of the external field, H_e for each mode depends on its direction, on the demagnetizing factors (i.e. the shape) of the specimen, and, of course, on the magnitude of the crystal anisotropy coefficients. In the general case the four modes would succeed each other as H_e increases, but in a number of cases of practical importance one or two of the last three modes may not occur.

In determining the state of magnetization for a given applied field the basic condition to be satisfied is that the energy of the crystal specimen as a whole is a minimum with respect to variation of the directions of magnetization (there being one direction for each phase present) and of the fractional volumes of each phase. In the work under review this condition is not applied directly (except in the single-vector, or mode IV case) but rather the consequential condition, already described, as to the direction of the resultant field when two or more phases are present. The fact that in an applied field a two- or three-phase state may have a lower energy than a one-phase state is essentially due to the demagnetizing field energy, equal to $+\frac{1}{2}NI^2$, where I is the resultant magnetization, always being smaller in the former state than in the latter, for which its value is $+\frac{1}{2}NI_0^2$. The smaller the relevant demagnetizing coefficients, N, the wider the range of applicability of

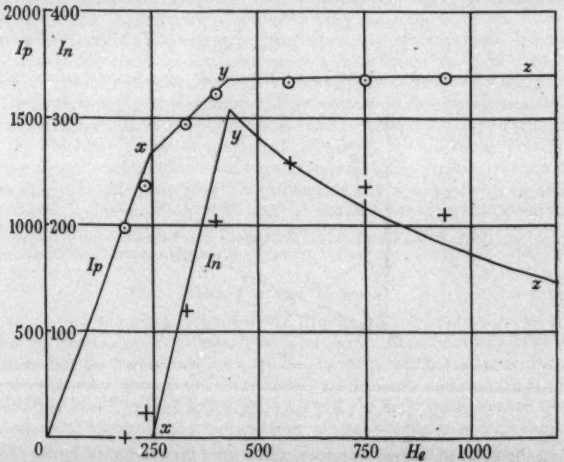


Figure 2.7. Calculated magnetization curves for an oblate spheroid of iron, equatorial plane (100). Applied field, H_e, at 20° to [100]. Demagnetizing coefficient, N, 0·184. Experimental points from Honda and Kaya (1926). Assumed values: $K_4=4{\cdot}2\times10^5$ erg cm^{-3}, $I_0=1720$ gauss. (Lawton and Stewart 1948.)

(*Reproduced, by kind permission, from the 'Proceedings of the Royal Society'.*)

the single vector treatment. Within a given mode, as the applied field is increased, the adjustment to the minimum energy state occurs physically by change of the directions of magnetization and change in the relative volumes of the phases. The reduction to zero of the fractional volume of any phase marks the transition to another mode.

In all but the very simplest cases (essentially those considered in § 2 (ii)) the calculation of the magnetization curves (i.e. of I_p and I_n as functions of H_e and θ) is troublesome and lengthy, and reference must be made to the original papers for details. In Figure 2.7, due to Lawton and Stewart, illustrative theoretical curves and experimental points are shown for a fairly simple case, that of a (100) spheroid,

with the angle, θ, between the applied field and the [100] direction equal to 20°. The parts Ox of the curves correspond to the mode I region (see Table 2.1) for which the resultant field, H, is zero. From this,

$$H_e - NI_p = 0, \qquad NI_n = 0,$$

giving

$$I_p = H_e/N, \qquad I_n = 0. \qquad \ldots\ldots (2.16)$$

The details of the changes in the relative volumes of the phases present during this stage are not required and are, in fact, indeterminate with the conditions assumed. The limit of the range occurs when there are two groups of domains (two phases) with magnetization directions along the two easy axis directions, [100], [010], nearest to the field, the resultant field still being zero. It may easily be shown that for the state corresponding to the 'knee' at x,

$$H_e = \frac{NI_0}{\cos\theta + \sin\theta}, \qquad I_p = \frac{I_0}{\cos\theta + \sin\theta}, \qquad I_n = 0, \qquad \ldots\ldots (2.17)$$

and that the volumes of the [100] and [010] phases are proportional to $\cos\theta$ and $\sin\theta$.

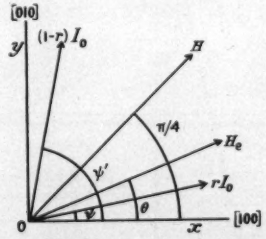


Figure 2.8. Magnetization of a (100) spheroid with the applied field in the (100) plane for the two-phase range (mode III).

H_e, applied field. H, resultant of applied and demagnetizing fields. r, fractional volume of phase with domains with I_0 at an angle ψ to the [100] axis.

At x the mode of magnetization changes from I (or a modified form of I, since there are effectively four rather than six phases) to III, with two phases. The directions of field and magnetization in this two-vector range (xy in Figure 2.7) are illustrated in Figure 2.8.

The resultant field, H, is symmetrical with respect to the easy axes (i.e. H is along [110]), and the angles between the magnetization vectors and the nearest easy axis are equal (i.e. $\psi + \psi' = \pi/2$). The variation of I_p and I_n with H_e, for a given θ, can be determined numerically from equations expressing the basic condition (that the energy is a minimum) or the consequential conditions, various short cuts being possible (e.g. by making use of the already known relation (2.12)). Physically, as H_e increases, the change occurs partly by rotation of the domain magnetization vectors, and partly by boundary movement processes resulting in a change in the relative volumes of the two phases. Referring to Figure 2.8, ψ increases and ψ' decreases, the sum remaining constant, and r increases from the value $\cos\theta$ (for $\psi = 0$) to unity. With the disappearance of the second phase ($1 - r = 0$), at the points y in Figure 2.7, mode III is succeeded by mode IV, and

the variation of I_p and I_n with H (yz in Figure 2.7) can be calculated by the single vector treatment. The good agreement between theory and experiment is clearly apparent in Figure 2.7.

Treatments of the magnetization of iron crystals in the form of (100) spheroids and of 'long rods' of particular crystal orientations are given both by Néel and by Lawton and Stewart, comparison being made with the experimental results of Webster (1925), Honda and Kaya (1926), Sizoo (1929) and Kaya (1933). In later work Lawton (1949) has considered in detail the case of the oblate spheroid with a (110) equatorial plane, and Stewart (1949) has extended the general treatment to cover the effects of tension. In most cases there is no significant difference between theory and experiment. That the experimental curves are usually more rounded than the theoretical ones is probably attributable to an effect of internal strains. The wider discrepancies occasionally found seem more likely to be due to experimental uncertainties (e.g. as to the composition, degree of homogeneity and freedom from internal strain, and precise crystal orientation of the specimen) than to any inherent defect in the theoretical treatment. Hysteresis effects and effects associated with domain boundary energies are deliberately left out of consideration. Tidier methods of developing the treatment may be evolved, and further applications and extensions of it will no doubt be made. It may be said, however, that in principle, as a result of the work outlined, the formal theory of the magnetic properties of ferromagnetic single crystals is now complete, in that the observable properties within its scope can be completely coordinated. The explanation of the magnitude and the temperature variation of the formal coefficients, the magneto-crystalline anisotropy coefficients, is a problem for atomic, as distinct from macroscopic, theory.

2(iv). *Atomic Theory*

The atomic and statistical theory of the magneto-crystalline anisotropy of metallic ferromagnetics is extremely complex, and no quantitatively satisfactory treatment has yet been developed. The papers which have appeared on this subject since 1934, though few in number, are concentrated in content and do not lend themselves to brief summary. All that can be done here is to indicate the methods of approach adopted and the general character of the conclusiohs reached.

Magnetic interaction. It is convenient to refer first to the work McKeehan (1933, 1937 b). After the slow and reluctant abandonment of the venerable hypothesis that the basic characteristics of ferromagnetics might be due to magnetic interactions between elementary carriers of the magnetic moment, appropriately shaped and disposed, it was still reasonable to suppose that these interactions might account for some of the secondary effects. McKeehan's work is essentially a detailed quantitative investigation of this possibility. Assuming that the elementary carriers are multipoles of moment $\mathbf{P}$, and that the multipole structure has axial and polar symmetry, the expression for half the mutual potential energy of a pair of multipoles with the moments parallel may be written

$$U = -(P^2F_2)(ar)^{-3}P_2\{\cos(\mathbf{P}_1,\mathbf{r})_1\} - (P^2F^4a^2)(ar)^{-5}P_4\{\cos(\mathbf{P}_1,\mathbf{r}_1)\} - (P^2F_6a^4)(ar)^{-7}P_6\{\cos(\mathbf{P}_1,\mathbf{r}_1)\}\ldots, \qquad \ldots\ldots(2.18)$$

where $\mathbf{P}_1$ and $\mathbf{r}_1$, are unit vectors in the directions of $\mathbf{P}$ and $\mathbf{r}$, whose numerical magnitudes are P and ar, a an appropriately chosen unit of length related to the

lattice constant, the F_r magnetic form factors, and the P_r zonal surface harmonics. The successive terms correspond to dipole, quadrupole, sextupole.... energy. The form factors may be calculated without undue difficulty for particular multipole structures, corresponding to specific distributions of magnetic moment. Distributions of sufficient generality are obtained by supposing the moment to be spread along two symmetrical parallels of latitude on a sphere of radius ρa, and directed parallel to the polar axis; the limiting cases correspond to an equatorial and polar distribution of magnetic moment. To find the dependence of the energy per unit volume on the direction of magnetization relative to the crystal axes, it is necessary to determine the sum of expressions of the form (2.18) over all pairs of magnets with one magnet in common. Extensive tables giving the results of the laborious evaluations of most of the required lattice sums are presented in McKeehan's earlier papers (1933). In the later paper (1937 b) further sums are evaluated, the form factor is introduced, and explicit calculations are made of the anisotropy coefficients for assemblies of multipoles of moment and lattice distribution in accordance with the experimental data for various ferromagnetic metals and alloys. (Only bare reference can be made here to McKeehan's subsequent extensions of his general methods of calculation to the treatment of homogeneously strained crystals of the ferromagnetic elements (1938 a), and of crystals of pyrrhotite and magnetite (1938 b).) The results depend, of course, on the structure assumed for the multipole; this is taken to correspond either to the equatorial or to the polar distribution of magnetic moment, as mentioned above, on a sphere of radius equal to that estimated for the mean radius of the 3d shell (Slater 1930). For cubic crystals there is no contribution to the anisotropy from the dipole interaction. For iron, assuming an equatorial distribution, the calculated value of K_4 (in 10^5 erg cm^{-3}) is $+255$ and of K_6 -1671. It may be recalled that the 'observed' value of K_4 at room temperature is about 4·7, and that the value at absolute zero is probably not much greater (see Figure 2.4). For nickel, assuming a polar distribution, the values obtained for K_4 and K_6 are -36 and $+287$, again much higher than the observed values. It may be noted that for both iron and nickel the sextupole part of K_4 is considerably greater than the quadrupole part. The difference in sign of K_4 for iron and nickel is a consequence of the assumed difference in the distribution of moment in the multipoles, an assumption which is to be regarded as a tentative suggestion to account for the observations. For hexagonal cobalt, the dipole contribution to the anisotropy term K_2' is appreciable, and of the right sign, but by itself far too small. Inclusion of the quadruple and sextupole contributions, with either extreme form of moment distribution, gives a total roughly of the order of magnitude required, but of the wrong sign. Values for the anisotropy coefficients calculated in the way outlined are given by McKeehan for many alloys as well as for the elements. It is to be noted that, throughout, the linear dimensions of the multipole structure are assigned in a non-arbitrary manner, and that the distribution of moment is assumed to be either of the extreme polar or equatorial form. With intermediate forms, and with arbitrarily adjusted dimensions, it would undoubtedly be possible to obtain agreement with the observed value of the leading anisotropy coefficient at a particular temperature, but this in itself would mean very little. McKeehan himself makes surprisingly little comment on the significance of the agreement (or disagreement) with experiment of the results of these extensive calculations. The general conclusion to be drawn, it may be suggested, is that the formal

characteristics of magneto-crystalline anisotropy are consistent with basic interactions formally similar to those between magnetic multipoles (implying, essentially, elementary carriers with a distribution of magnetic moment over a space of atomic dimensions); but that, quite apart from the fact that temperature effects are entirely unaccounted for, the type of structure required to give agreement with experiment even as to order of magnitude is incompatible with what is now known about the effective magnetic 'elements' in ferromagnetic metals. McKeehan's papers, in short, do not give a satisfactory atomic theory of magneto-crystalline anisotropy; but they contribute indirectly towards such a theory, and the precise and extensive numerical results for the energies and internal fields associated with the purely magnetic interactions in arrays of multipoles have a permanent value in this and other connections.

Spin–orbit and orbit–orbit interaction. The problem of magneto-crystalline anisotropy may be approached in a physically more satisfying manner by considering first what is known, from other evidence, about the elementary carriers of the magnetic moment. For the ferromagnetic metals experiments on the gyromagnetic effect give a g value equal to, or only a few per cent less than, 2, the value corresponding to electron spins as carriers. The orbital contribution to the magnetization, though not necessarily zero, can hardly in metals be as much as one tenth of the whole. Now in cubic crystals the purely magnetic dipole interaction of the spins gives no anisotropy, and quite generally the magnetic interaction is far too small to account for the observed effects. The very much larger exchange interaction which is responsible for the 'molecular field' in ferromagnetics may be expressed as formally equivalent to a coupling between spins, but this, depending on the angle between two spins but not on the direction of the line joining them, does not give rise to anisotropy. It is now usually accepted that the anisotropy arises in some way from spin–orbit coupling, a view apparently first suggested by Bloch and Gentile (1931), and developed in detail by Van Vleck (1937). Van Vleck gives a long discussion of the whole problem, with full references to earlier work. The almost complete quenching of an orbital moment effect in the magnetization of metals may be attributed to a strong orbit–orbit coupling between different atoms of anti-ferromagnetic sign. Alternatively expressed, the energy distribution of orbital states (conditioned by, and itself conditioning, the crystal symmetry and lattice dimensions) is such that the occupational distribution is little modified by changes in direction of the resultant local magnetization arising from the spins. There is, however, some degree of spin–orbit coupling, with the result that the orbital distribution will vary to some extent with the direction of the magnetization relative to the crystal axes, and, in consequence, the orbit–orbit coupling between different atoms will change. This orbit–orbit coupling is not thought of as being primarily magnetic in character, but rather of the 'orbital valence' type. An exact treatment of the perturbation problem which arises is virtually impossible. Van Vleck shows, however, that the Hamiltonian function containing the formally correct terms corresponding to the spin–orbit and the orbital valence effects may be replaced, with good approximation, by a Hamiltonian containing terms of the dipole and quadrupole type with appropriate coupling coefficients. Even with this simplified Hamiltonian, the problem remains very complex, and in the second part of his paper Van Vleck develops the elaborate mathematical machinery required to deal with it. In view of the uncertainties in the detailed formulation of the problem, and its inherent difficulty, it is not

surprising that no precise numerical results emerge. Little more can be said than that the experimental anisotropy coefficients are consistent as to order of magnitude (to within a factor of 10, say) with such rough estimates as can be made of the various coupling coefficients. This also applies to the contemporary work of van Pepye (1938), who treats a limited aspect of the problem more rigorously than Van Vleck, and to the later work of Brooks (1940), who bases his mathematical treatment on the 'itinerant' or collective electron approximation (I, pp. 69–75) rather than on the Heitler–London approximation (I, pp. 53–58, 66–68).

The theoretical results for the temperature variation of K_4 as given by Van Vleck (1937, pp. 1194, 1196) show rough qualitative agreement with experiment for iron, for which a quadrupole 'model' is used. With the dipole model used for nickel (which would give the wrong sign of K_4 for iron), the calculated temperature variation is widely different from that observed in that there is no indication of the large increase in the magnitude of K_4 below room temperature (see Figure 2.4). An expression derived by Akulov (1936) indicating a variation of K_4 with temperature approximately as I_{0T}^{10}, where I_{0T} is the saturation magnetization at temperature T, has often been quoted as giving good agreement for iron over a fairly wide temperature range. The premises of Akulov's theoretical treatment are, however, so artificial that any such agreement can only be regarded as fortuitous; moreover, even if the premises were valid, the power relation would be a good approximation only for values of $(I_{00}-I_{0T})/I_{00}$ much less than unity. The temperature dependence of the anisotropy of cobalt has been treated theoretically by Vonsovski (1940), who follows the general method of Bloch and Gentile. An expression is obtained of the form

$$K_2' = (K_2')_0(1+2e^{-a/T}-3e^{-b/T}), \qquad \ldots\ldots(2.19)$$

which fits the observations reasonably well, including the change of sign of K_2' (see Figure 2.4), with values for a and b which are of the same order of magnitude as the very rough theoretical estimates.

In view of the central importance of magneto-crystalline anisotropy in connection with the general behaviour of ferromagnetics in low and moderate fields it is unfortunate that the atomic theory of the effect, as developed up to the present, is quantitatively so imperfect. In dealing with the problem, however, the starting point is necessarily some theoretical scheme for the basic electronic structure of the metal, and a (preferably consistent) theoretical scheme for the effects of exchange interaction; and when the quantitative deficiencies of all such schemes are borne in mind it is somewhat remarkable that any theoretical treatment of further complicating effects can lead to results in even the roughest quantitative agreement with observation. It may be hoped that different methods of approach to the various problems may lead to simpler theoretical coordinating schemes by which the essential significance of the diversity in magnitude, sign and temperature variation of the anisotropy coefficients of different materials may be more clearly understood. This, however, is perhaps over-optimistic, and it may be well to conclude this section with a quotation from the published version (1947) of some lectures given in Paris in 1939 by Van Vleck, who has made the most determined and elaborate attack on this problem of magneto-crystalline anisotropy: 'Le problème est tellement compliqué que des résultats quantitatifs satisfaisants ne pourront être obtenus, si jamais on les obtient, que dans un avenir très éloigné.'

§ 3. DOMAIN STRUCTURE

3 (i). *Theoretical and General*

In this section consideration is given to experimental and theoretical work which bears directly on the question of the size, shape and distribution of domains in ferromagnetics. With the acceptance of a basic explanation of ferromagnetism in terms of exchange interaction giving rise to molecular field effects (I, pp. 65–78), the behaviour of ferromagnetics in low and moderate fields provides indirect but irrefutable evidence of a domain 'structure' of some kind. It does not, however, enable the character of that structure to be specified in any detail. The hope that the study of Bitter patterns (§ 3 (ii)), of the Barkhausen and related effects (§ 3 (iii)), and of small particles and thin films (§ 3 (iv)) might yield more precise and unambiguous information was long deferred (cf. I, p. 65), and is only now being fulfilled. Attention must here be devoted mainly to a few of the more recent papers, and it is not possible to trace the course of development in detail. It should be understood, however, that the very rapid advance which has recently been made is due not so much to the introduction of essentially novel theoretical ideas or experimental methods as to the effective following up of earlier investigations along different lines which have converged on the domain structure problem. For general reviews of the subject matter of this section, reference may be made to the books of Bitter (1937) and of Becker and Döring (1939), to papers by Brown (1940 a, 1945), and by Kittel (1949).

The term 'domain' has been used in more than one sense (cf. Stoner 1938). It will be used here not of a fixed region of the specimen, free from heterogeneities of composition or of strain, but in the more usual sense of a bounded region throughout which the magnetization is uniform, except, possibly, for slight fluctuations in the direction or the intensity. The shape and size (and consequently the number) of domains in a specimen changes with change in magnetization of the specimen as a whole. The domain structure is not therefore static, nor is it, in general, uniquely dependent on the macroscopic state of magnetization of the specimen. When the specimen is saturated or, more generally, in a one-phase state (see § 2 (iii)), the whole specimen constitutes a single domain. The direction of magnetization of a domain depends on the magneto-crystalline and strain anisotropy, and on the effective field. When strain effects are negligible, and the effective field is small, a domain is magnetized along one or other of the equivalent easy axis directions; in iron along a cube edge, in nickel a cube diagonal, and in cobalt along the hexagonal axis. Within a single crystal region of an iron-like specimen, the directions of magnetization in contiguous domains, under the conditions mentioned, may therefore be at angles of 90° or 180°, and the transition zones between them are referred to as 90° or 180° boundaries, or walls. (In nickel-like crystals, under the same conditions, the possible angles are 70·53°, 109·47°, and 180°; in cobalt-like crystals, only 180° walls could occur.) As such extensive reference is made to 90° and 180° walls, and they are usually considered in numerical calculations, the erroneous impression is liable to be gained that (at least in iron-like crystals) boundaries are invariably of one of these two types. Brief consideration of the course of magnetization of single crystals, as already described (§ 2(iii)) shows, however, that 180° boundaries are present only in the first mode of magnetization of cubic crystals (when more than three phases are present). . . Further, in the three- and two-phase modes, the angle between domain

magnetization directions, initially 90°, gradually decreases as the resultant magnetization increases (cf. Figure 2.8). Néel (1944 b, c) gives some numerical results showing how the boundary energy decreases to zero as the domain magnetization vectors approach parallelism. The obvious problem raised by grain boundaries in polycrystalline materials has not yet been considered in any detail in the present connection.

Boundary energy. Domain boundaries are sometimes referred to as Bloch walls, the first rough estimates of the energy associated with them and of their effective thickness having been made by Bloch in 1932. Bloch's general treatment has since been refined and extended, notably by Landau and Lifshitz (1935) and by Néel (1944 c). In the following account, based mainly on their papers, the results of some supplementary calculations have been included.

A characteristic of domain boundaries, a full recognition of the importance of which underlies much of Néel's work (1944 a, b, c), is that, in general, the normal component of the magnetization has the same value not only on both sides of the boundary, but also in the transition region. This is because the development of internal free poles (corresponding to a non-zero value of div **I**) would usually involve an increase of energy very much greater than in some other possible arrangement in which free pole formation does not occur. Thus if the magnetization vector, $\mathbf{I}_0$, makes an angle θ with the normal to the boundary, and its projection on the boundary surface an angle ϕ with some selected zero direction, the change of direction of $\mathbf{I}_0$ in passing through the boundary occurs by a gradual rotation of $\mathbf{I}_0$ about the normal, θ remaining constant, and ϕ changing from, say, ϕ_1 to ϕ_2. For a 180° boundary, for example, θ is necessarily 90°, and ϕ changes from ϕ_1 to $\phi_1 \pm 180°$.

Taking the 180° boundary in illustration, suppose that the boundary is parallel to the yz plane, the normal being along x, and that an easy direction of magnetization of the crystal (e.g. a cube edge for unstrained iron) lies along the z axis; further, that in passing through the boundary the direction of magnetization changes from $+z$ to $-z$. The boundary energy arises from both exchange interaction and magneto-crystalline or strain anisotropy. The exchange energy is proportional to $1-\cos\delta\phi$, where $\delta\phi$ is the angle between the spin directions in neighbouring layers. The contribution per unit volume may therefore be expressed in the form

$$F_e = \alpha a^2 (d\phi/dx)^2, \qquad \ldots\ldots (3.1)$$

where a is a lattice constant, and α is proportional to the exchange energy per unit volume (in terms of the Weiss molecular field treatment to $\frac{1}{2}N_W I_0^2$), with a numerical factor, usually of the order 10^{-1}, depending on the lattice type and the number of effective spins per atom. The anisotropy energy per unit volume may be written

$$F_a = \beta f(\theta, \phi), \qquad \ldots\ldots (3.2)$$

where β is a non-particularized anisotropy coefficient. The energy and effective thickness of the boundary are determined from the condition that the sum of the energies F_e and F_a, integrated over the x range in which ϕ changes from 0° to 180°, shall be a minimum.

There is usually no difficulty in determining the form of $f(\theta, \phi)$ for the type of boundary under consideration. The general method may be illustrated by putting

$$F_a = \beta \sin^2 \phi. \qquad \ldots\ldots (3.3)$$

This relation holds, at least approximately, if the z direction is the easy direction in an unstrained cobalt-like crystal (when $\beta = K_2'$), or if it is the direction of application of tension (compression) in a material of positive (negative) isotropic magnetostriction (when $\beta = 3\lambda z/2$, where λ is the saturation magnetostriction coefficient and σ the tension), provided it can be assumed that the natural magneto-crystalline effect is relatively small. The energy per unit area of the domain boundary, usually denoted by γ, is then given by

$$\gamma = \int \{\alpha a^2 (d\phi/dx)^2 + \beta \sin^2 \phi\}\, dx, \qquad \ldots\ldots (3.4)$$

where the integral extends over the x range in which ϕ changes from ϕ_1 to $\phi_1 + \pi$. A straightforward variational treatment then gives, as a condition that the integral shall have a stationary value,

$$\alpha a^2 \frac{d}{dx}\left(\frac{d\phi}{dx}\right)^2 = \beta \frac{d}{dx}(\sin^2 \phi). \qquad \ldots\ldots (3.5)$$

Integrating this equation,

$$\alpha a^2 \left(\frac{d\phi}{dx}\right)^2 = \beta \sin^2 \phi + c. \qquad \ldots\ldots (3.6)$$

For $|x| \to \infty$, $\phi \to 0$ or π, and $d\phi/dx \to 0$, and the integration constant is therefore zero. The two terms in the integrand of (3.4) are therefore equal, i.e. the exchange and anisotropy contributions to the energy are equal. It is convenient to introduce

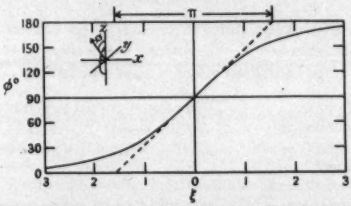


Figure 3.1. Variation of direction of magnetization in a 180° domain boundary in a specimen with a single easy axis of magnetization.

x, distance along normal to boundary. ϕ, angle between I_0 and easy axis in the yz plane. $\xi = x/b_0$. $b_0 = (\alpha a^2/\beta)^{1/2}$. α, β, exchange and anisotropy energy density coefficients. a, lattice constant.

a unit of length, b_0, normal to the boundary, and a unit, γ_0, for the energy per unit area of the boundary, given by

$$b_0 = (\alpha a^2/\beta)^{1/2}, \qquad \gamma_0 = (\alpha a^2 \beta)^{1/2} = \beta b_0 = \alpha a^2/b_0, \qquad \ldots\ldots (3.7)$$

and to use a non-dimensional parameter, ξ, for length measured normal to the boundary (i.e. along x), given by

$$\xi = (x/b_0). \qquad \ldots\ldots (3.8)$$

Putting $\phi = \pi/2$ at $\xi = 0$, integration of (3.6) yields the relation between ϕ and ξ in the forms

$$\left.\begin{aligned} \cos\phi &= -\tanh \xi, \qquad \sin\phi = \operatorname{sech} \xi, \\ \phi &= 2\tan^{-1} e^{\xi}, \qquad \text{or } \xi = \ln \tan(\phi/2). \end{aligned}\right\} \qquad \ldots\ldots (3.9)$$

The character of the transition from $\phi = 0°$ to $\phi = 180°$ is shown in Figure 3.1.

It will be clear that the boundary, if it satisfies the minimum energy condition, has no sharply defined width, but as shown in the figure, the effective width in this case may be conveniently taken as πb_0 (approximately 0·74 of the total change in ϕ takes place over this distance). The energy per unit area of the boundary is readily found by integration as

$$\gamma = 4\gamma_0. \qquad \ldots\ldots (3.10)$$

The last expression for γ_0 in (3.7) shows at once how the energy arising from departure from parallelism of the spins is reduced by the gradualness of the transition, for, with an abrupt change of direction of I_0, γ_0 would be of the order αa. It may be noted (see further below) that b_0 is usually of the order 10^{-6} cm., and γ_0 of the order 1 erg cm^{-2}.

The method of treatment of domain boundary energy and thickness outlined above for the 'uniaxial' case is generally applicable, but the problem is naturally more troublesome when there is more than one easy axis and the required anisotropy energy expression (3.2) has to be derived from a basic expression of the type (2.1). Néel (1944 c) has considered in detail the energy of boundaries in an iron crystal, and some of his results are shown in Table 3.1. In all cases the change is from one easy direction (cube edge) to another, the normal component of the magnetization remaining constant.

Table 3.1

Relative Energies of Domain Boundaries in Iron (as calculated by Néel 1944 c)
The values tabulated are of γ/γ_0, where γ is the energy per unit area, and the estimated value of γ_0 is 0·7 erg cm^{-2}.

Type of boundary	Direction of normal to boundary		
	[100]	[110]	[111]
90°	1·0	1·727	1·185
180°	2·0	2·760	—

The character of the transition in the 90° boundaries is similar to that shown in Figure 3.1. For the 180° boundaries, however, a difficulty arises. A 180° boundary parallel to a (100) plane might be regarded as two successive 90° boundaries; but the minimum energy condition would require an infinite separation between them. Néel shows that this difficulty is overcome when magnetostrictive anisotropy energy is also taken into account. This gives rise, in the general case, to an effective attraction between neighbouring 90° boundaries, and in an ideal crystal of iron, in which boundaries can move freely, the only stable 90° boundaries would be those normal to a binary axis. Although magnetostrictive energy is an important factor in determining the effective width and stability of boundaries, its effect on the overall boundary energy is negligible. Graphs plotted from Néel's formulae for the variation of direction of I_0 in 90° and 180° boundaries normal to a [100] direction are shown in Figure 3.2.

It will be clear from the above account that by the methods now available reliable theoretical estimates may be made of the relative energies of boundaries of different types in any particular material whose magnetic, magneto-crystalline and (more particularly if material under stress is to be considered) magnetostrictive characteristics are known. Estimates of the absolute values of the energy per unit area, and of the relative energies of boundaries in different materials, are much

more uncertain, mainly owing to the many uncertainties in the detailed treatment of exchange interaction effects (cf. I, pp. 65–78). It is usually assumed in the calculations that have been made that the interaction is effectively between nearest neighbours, and proportional to the cosine of the angle between the spins, and that the difference in energy per unit volume between the states of random and parallel spin orientation is $\frac{1}{2}N_W I_0^2$, N_W being the Weiss molecular field coefficient. It is then not difficult to show (cf. Landau and Lifshitz 1935) that for a simple cubic lattice, with lattice constant a, the exchange energy per unit volume for a transition region in which the magnetization vectors are normal to a cube axis, say the x axis, and vary in direction only with x (cf. (3.1)) is given by

$$F_e = \tfrac{1}{6}(\tfrac{1}{2}N_W I_0^2)a^2(d\phi/dx)^2. \qquad \ldots\ldots(3.11)$$

A generalization of this expression is given by Stoner and Wohlfarth (1948, p. 634). By using, rather than the lattice constant a, a parameter a_1, where a_1^3 is the volume per atom, the same numerical factor, to within a few per cent, is obtained for both face-centred and body-centred cubic lattices. The coefficient N_W is

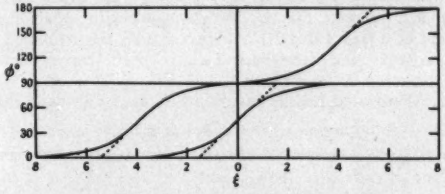


Figure 3.2. Variation of direction of magnetization in 90° and 180° boundaries normal to a [100] direction in iron. (From formulae given by Néel 1944 c.)
$\xi = x/b_0$. x, distance normal to boundary. Estimated value of b_0 (Néel), 165×10^{-8} cm.

expressed in terms of experimentally determinable quantities, including the Curie temperature, θ, using the modified molecular field treatment (I, pp. 55–58), for this purpose a reasonable approximation. Further, explicit allowance is made for the number, q, of effective spins per atom (2·22 for iron, 0·60$_8$ for nickel, see I, Table 4.1, p. 86) in what appears to be the most plausible way consistent with present knowledge. When the intensity of magnetization, I_{0T}, is approximately equal to that at absolute zero, I_{00}, the final expression (derived from equation (6.6), *l.c.*, which, together with (6.5), should be corrected by the insertion of a multiplying factor q) takes the very simple form

$$F_e = \tfrac{1}{10}q^{1/3}\frac{k\theta}{a_1}\left(\frac{d\phi}{dx}\right)^2 = \alpha_1 a_1^2\left(\frac{d\phi}{dx}\right)^2. \qquad \ldots\ldots(3.12)$$

By substituting $\alpha_1 a_1^2$ for αa^2 in (3.7), and using for β the values of K_4 for iron and nickel, and of K_2' for cobalt, approximate estimates of the unit lengths, b_0, and unit energies, γ_0, may readily be obtained. These are given in Table 3.2.

No rigour would be claimed for any of the expressions which have been used for F_e. In the light of the collective electron treatment of ferromagnetism it is believed that (3.12) almost certainly gives an underestimate, possibly by a factor of 2, but that it can be used with reasonable confidence for comparative purposes. It may be noted that the values of γ_0 and b_0 for iron in Table 3.2 agree closely with

those given by Néel (0·7 and 165 respectively), but this agreement is partly fortuitous as there are significant differences in the methods of derivation. In a method given by Lifshitz (1944) a quantity corresponding to $\alpha_1 a_1^2$ in Table 3.2 is evaluated from the coefficient of $T^{3/2}$ in Bloch's law of approach to saturation at absolute zero (I, p. 66). Bloch's spin-wave treatment is on a Heitler–London–Heisenberg basis, and the $T^{3/2}$ law is a rigorous solution of the formal problem which is considered, or, at least, a first approximation, valid in fact over only a very narrow temperature range, to such a solution. It has been shown by Wohlfarth (1949), however, that the conditions presupposed do not, in general, have a sufficiently close correspondence to those in actual ferromagnetic metals for the treatment to be applicable to them. The values of γ_0 to which the results of Lifshitz lead would be some 50% higher than those in Table 3.2, values which, as already stated, are almost certainly underestimates. Even if experimental values for boundary energies are eventually found to be in agreement with estimates made by Lifshitz, the theoretical method would require much fuller justification than it has so far received.

The most interesting point in Table 3.2 is the wide difference in the values for the different elements. For the 180° boundaries in cobalt (see (3.10) and Figure 3.1) and iron (see Table 3.1 and Figure 3.2), the estimated values of γ_0 are 8·2 and 1·2 erg cm^{-2}, and of the effective widths (in 10^{-8} cm.) about 160 and 1,400

Table 3.2. Estimated Energy and Width Units for Domain Boundaries

$$\gamma_0 = (\alpha_1 a_1^2 \beta)^{1/2}, \qquad b_0 = (\alpha_1 a_1^2/\beta)^{1/2}$$

	$\alpha_1 a_1^2$ 10^{-7} erg cm^{-1}	β 10^8 erg cm^{-3}	γ_0 erg cm^{-2}	b_0 10^{-8} cm.
Fe	8·3	4·7	0·62	130
Ni	3·4	0·51	0·14	260
Co	10·3	41	2·05	50

respectively. The indication is that for nickel the energy would be about a quarter, and the effective width twice as great as that for iron. These comparisons have a bearing on the domain sizes in the different elements, and on the character of powder pattern lines.

Domain shapes and sizes. The first soundly based attempt to estimate domain shapes and sizes quantitatively is that of Landau and Lifshitz (1935), who considered the particular case of a crystal with a single easy axis of magnetization. The principles implicit in their treatment have been concisely and very clearly discussed by Elmore (1938 a). The method has been developed further by Lifshitz (1944), and applied to iron as well as cobalt. The most detailed discussion of the whole problem of the magnetization curves, and the subdivision into domains of single crystals, more particularly of iron, is that of Néel. The first part of this work (1944 a), on the different modes of magnetization of a single crystal has already been considered (§ 2(iii)). The second and third parts (1944 b) deal with the shape and orientation of the domains, and with the secondary superficial domain structure and the thickness of domain layers.

The general principles can be brought out by consideration of a cobalt-like crystal in a demagnetized state in zero applied field. The domain structure proposed by Landau and Lifshitz is shown diagrammatically in Figure 3.3. The condition determining the dimensions of the domain pattern is that the free

energy of the crystal as a whole (or more conveniently, the average energy density) shall be a minimum. The direction of magnetization may differ in different domains, but the intensity is effectively constant, and equal, sufficiently closely, to the saturation magnetization, I_0. The main contributions to the energy, additional to that associated directly with the intrinsic magnetization in the domains (i.e. the exchange energy, which can be represented by $-\frac{1}{2}N_W I_0^2$, where N_W is the Weiss molecular field coefficient), arise from (*a*) the field energy, F_H, (*b*) the domain boundary energy, F_γ, and (*c*) the anisotropy energy, F_a. Symbolically, the condition is that F shall be a minimum, where

$$F = F_H + F_\gamma + F_a. \qquad \ldots\ldots(3.13)$$

The field energy, in general, arises from both the external and the internal fields, H_e and H_i respectively, and, on the understanding that energies are to be integrated over the volume of the crystal and averaged, may be written

$$F_H = F(H_i) + F(H_e) = -\tfrac{1}{2}\mathbf{H}_i \cdot \mathbf{I}_0 - \mathbf{H}_e \cdot \mathbf{I}_0. \qquad \ldots\ldots(3.14)$$

For the demagnetized state (or when the applied field is zero) the contribution from the last term is zero. The internal magnetic field, H_i, is derivable from a potential given by Poisson's equation,

$$\Omega = \int \frac{\mathbf{I} \cdot \mathbf{n}_1}{r}\, dS - \int \frac{\operatorname{div} \mathbf{I}}{r}\, dv. \qquad \ldots\ldots(3.15)$$

The internal field energy (which is always positive or zero) arises from the presence of 'free poles', either superficial ($\mathbf{I} \cdot \mathbf{n}_1 \neq 0$) or internal ($\operatorname{div} \mathbf{I} \neq 0$). The energy

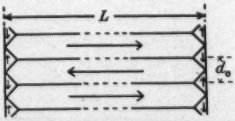


Figure 3.3. Domain structure in a magnetically uniaxial crystal in the demagnetized state. (*After* Landau and Lifshitz 1935.)
The easy axis is horizontal in the plane of the diagram. The domain boundaries are perpendicular to this plane. The arrows show the directions of magnetization.

associated with the presence of free poles is usually relatively large, and, except for the free pole distribution necessarily associated with the macroscopic state of magnetization of the specimen, any arrangement satisfying the minimum energy condition will be one in which, to a close approximation, free poles do not occur. The bearing of this on the character of the transition in domain boundaries has already been discussed. An important corollary, pointed out by Néel, is that domain boundaries must be plane or nearly plane. In particular, corrugated boundaries, represented in cross section by wavy lines, are precluded. In the case illustrated in Figure 3.3, it is clear that a high energy would result from superficial free poles if the primary domains extended right up to the surface. This energy may be reduced by the formation of superficial 'closure domains' of the type shown, triangular prisms normal to the plane of the diagram. With this arrangement $F(H_i)$ in (3.14) is zero, and consequently also F_H in (3.13). The energy to be minimized is thus the sum of the domain boundary

energy, and the anisotropy energy due to the deviation of the magnetization in the prisms from an easy direction. Taking the form (3.3) for the anisotropy energy, and using γ as above for the energy per unit area of the primary boundaries (the contribution from the prism boundaries may be neglected), it is easily found that the energy per unit volume is given by

$$F = \frac{\gamma}{d} + \tfrac{1}{2}\frac{d}{L}\beta, \qquad \ldots\ldots(3.16)$$

where d is the thickness of the domain layers and L the length of the crystal in the easy axis direction (see Figure 3.3). The energy is a minimum for

$$d = d_0 = (2L\gamma/\beta)^{\frac{1}{2}}, \qquad \ldots\ldots(3.17)$$

when

$$F_{\min} = (2\beta\gamma/L)^{\frac{1}{2}}. \qquad \ldots\ldots(3.18)$$

A comparison of 'theoretical' and experimental values of d_0, as obtained from patterns, is made below (§3(ii)), but it may be mentioned here that, using the values of γ_0 and β in Table 3.2, the value estimated from (3.17), with $L = 1$ cm., for d_0 in cobalt ($\gamma = 4\gamma_0$) is about 2×10^{-3} cm., which is at least of the order of

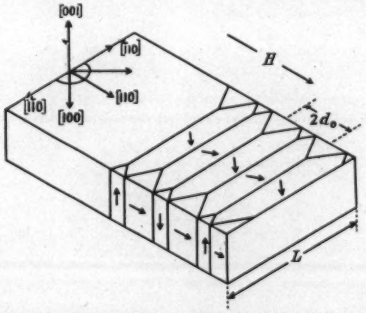


Figure 3.4. Domain structure in a monocrystalline iron strip, with field and magnetization along a [110] direction.
(*After* Néel 1944 b.)

magnitude observed. It is noteworthy that, with the proposed model, the thickness of the domain layers depends explicitly on the length of the crystal parallel to the easy axis.

The domain structure problem is more complicated for cubic crystals with three equivalent easy axes, but Néel shows that in the two- and three-phase modes of magnetization in iron the primary domains are again of the layer type, the domains being piled on top of each other like the leaves (*feuillets*) of a book. He considers in detail the case of a monocrystalline strip or slab with the field and resultant magnetization in a [110] direction. The proposed domain structure is shown diagrammatically in Figure 3.4.

The diagram applies to the two-phase range of magnetization between remanence and saturation. At remanence the magnetization vectors on the (100) surface are alternately along [100] and [010] directions, but as the magnetization

increases they turn towards the [110] direction. At the same time there is a change in the relative sizes of the two types of prismatic closure domain on the (110) and ($1\bar{1}0$) surfaces, corresponding to a preferential growth of those magnetized parallel to the field. It is shown that the magnetization in these prismatic domains is either along the field direction or very nearly perpendicular to it. The domain width, d_0, is calculated by a method similar to that used in deriving (3.17), but more complicated, for both the anisotropy energy and the boundary energy depend on the resultant magnetization. Néel's final result is that d_0, proportional to $L^{1/2}$ as before, decreases from a high value near the remanence point, for which $I/I_0 = 0{\cdot}707$, to a value, for $L = 1$ cm., between 65 and 70×10^{-4} cm. for $I/I_0 = 0{\cdot}85$ (corresponding to an effective field of about 190 oersted). A diagram is given showing that the estimated theoretical curve passes reasonably well through the somewhat scattered values obtained from observations on powder patterns by Kaya (1934 a, b).

It is, of course, not possible to enter here into the details of the theoretical derivation of even the relatively simple structure illustrated in Figure 3.4, or to consider more complex cases which have been discussed to some extent by both Néel and Lifshitz. Essentially, once the general character of the domain pattern has been established, the scale of the pattern is determined by a balance between the boundary energy of the primary domains and the anisotropy energy (including a sometimes not negligible magneto-elastic contribution) of the closure domains. The relation (3.17) sufficiently indicates the general orders of magnitude. It must be remembered that as regards absolute dimensions, little more than order of magnitude estimates can at present be made. This is due partly to the possible complexity of the superficial closure domain structure, partly to the uncertain values of the anisotropy and related coefficients (cf. §2(ii)), but mainly to the uncertainties in estimating the energies associated with exchange interaction in domain boundaries.

There are a number of special matters which have been examined theoretically which are conveniently deferred for consideration in connection with experimental results. These include the simple and multiple branching of domain layers, under certain conditions, as a surface is approached (Lifshitz 1944), the formation of special domain structures round holes or non-ferromagnetic inclusions to reduce free pole energy (Néel 1944 d), the behaviour of large domains in stressed material (Döring 1938, Haake 1939), and formation or non-formation of domain boundaries in thin films and small particles (Kittel 1946, and others).

It may be useful, in closing this theoretical section, to draw attention briefly to a few general points in the treatment of the domain structure problem. Running through the inter-related series of investigations, as presented here, is a search for states of minimum energy. These states can, in principle at least, be described with increasing precision and detail as effects corresponding to successively decreasing energy changes are taken into account. In the first approximation, considering only exchange interaction, the largest energy effect, a single crystal would be magnetized to the saturation value appropriate to the temperature. Next, the effects of an applied field and crystalline or strain anisotropy are considered, and also, in the Néel–Lawton–Stewart work (§2(iii)) the effect of the demagnetizing field. The at first sight remarkable result of their investigations is that the number of 'phases', each associated with a particular direction of magnetization, which can co-exist in a single crystal is very limited, not more than

six for cubic crystals, and, in general, over most of the magnetization range, only two or three. For given conditions, the magnetic characteristics (i.e. the directions of magnetization) and the relative volumes of these phases can be calculated (and hence the change of bulk magnetization with change of field), not perhaps easily, but, in most of the physically more important cases, without undue difficulty. This greatly facilitates the consideration of domain structure, for the starting point of the treatment can be taken as the already known number of domain types (as specified by their directions of magnetization) under the given conditions, and, to very close approximation, the proportions of the volume which they occupy. At this stage, to determine the actual spatial distribution of domains, other energy effects, in general smaller, must be taken into account, as already described. The complex problem of the enhanced exchange energy associated with the deviation from parallelism of the spins can be isolated in the boundary energy problem, and the calculated boundary energies can then be utilized, without further direct reference to exchange interaction, in determining the domain boundary pattern consistent with minimum energy. It should perhaps be stated explicitly that the pattern cannot be determined unambiguously merely by the application of a minimum energy criterion. This criterion fixes the scale of any given pattern, and enables a choice to be made between such patterns as may be devised. In devising patterns the most important working rule is perhaps that, for a first approximation, the arrangement should be such as not to give rise to internal or superficial free poles other than those necessarily associated with the bulk state of magnetization.

Finally, brief comment is perhaps desirable on a possible objection to domain theory which has been made by Brown (1945) in an interesting paper on the virtues and weaknesses of the domain concept, to the effect that although the theory may explain a great deal once domains are assumed, it does not explain their formation. He considers a single crystal specimen in the form of an ellipsoid with a direction of easy magnetization along a principal axis, initially saturated in this direction by a large applied field. He then shows that any small deviation of I_0 from the easy direction will increase the energy, and that as the field decreases no change will occur until the field becomes negative, and at least of the order of magnitude K/I_0, where K is an anisotropy coefficient. The magnetization would then reverse completely at this relatively high value of the field, and no subdivision into small domains (such as that shown in Figure 3.3) would occur at any stage. A tentative answer to this implicit objection to domain theory is that under the conditions assumed the course of the change with changing field would be as described. For an ideal ellipsoidal single crystal, perfectly uniform, no domain boundaries would be formed, in spite of the fact that in low fields, in zero applied field in particular, the saturated state is metastable, with a higher energy than for a state of low resultant magnetization, with domains magnetized in opposite directions. Any actual crystal, however, is imperfect, and as the field is reduced, domain boundary growth can take place, starting at holes, cracks or inclusions. The energy required to start the boundary formation is provided, so to speak, by the free pole energy. The extension of boundaries, once started, does not present so formidable a problem; it will be possible so long as it is accompanied by a continuous decrease in energy. There is no suggestion in domain theory that a ferromagnetic always necessarily takes up a domain configuration corresponding to the lowest possible energy, but the

possibility that a perfect crystal, in which domain boundary movement would be least restricted, might not form domains at all is at first sight somewhat startling. It may be suggested, however, that even the perfect ellipsoidal crystal might be persuaded to behave somewhat more normally by a few scratches with a pin.

3 (ii). *Powder Patterns*

General. Cracks and flaws in large ferromagnetic samples may be detected by dusting relatively coarse ferromagnetic powder over the surface. An analogous method, equally depending on the attraction of ferromagnetic particles from weaker to stronger fields, was introduced independently by Bitter and by von Hamos and Thiessen in 1931 for the examination of fine scale magnetic structure. Since then many investigations of powder patterns have been made, and the technique has been greatly refined. Powders have been used dry with some success (e.g. carbonyl iron powder), but for most purposes colloidal suspensions are more convenient. With a size greater than about 1μ, the particles tend to settle out; with a size less than about $0{\cdot}1\mu$, Brownian movement becomes prominent, and the particles, though they still concentrate satisfactorily in the stronger field regions if they are not too small, remain mobile, and the changing of a pattern with changing field can be readily observed (McKeehan and Elmore 1934 a, b). In recent years the most widely used 'powder' has been a colloidal suspension of magnetite Fe_3O_4, protected by soap (Elmore 1938 b). Clear patterns with recognizable regularities have been observed with unstrained material only on surfaces nearly coincident with simple crystallographic planes. The production of single crystal specimens, or at least of large-grained samples, is therefore an essential preliminary if significant results are to be obtained; the careful grinding of surfaces in required crystallographic planes, and the optical or x-ray determination of orientations are incidental techniques. Finally the surfaces must be suitably prepared, a mirror-like smoothness being desirable. With surfaces prepared by metallurgical polishing, McKeehan and Elmore (1934 a, b, 1936) obtained many intriguing maze-like patterns, some of which have often been reproduced. These patterns usually show, individually, a remarkable regularity of spacing, and this, together with their behaviour in changing vertical and horizontal fields suggested the presence of a stable block structure, possibly superficial, in the crystal. Later work, particularly by Elmore (1937), showed conclusively that any such structure must originate in the polishing operation; the maze patterns do not occur on smooth unpolished crystalline strips, but they are readily obtained in the region of broad shallow scratches made on them. To produce mirror-like surfaces free from strains, Elmore finally adapted and developed the method of electrolytic polishing (1939, 1940 a), a method which he applied successfully to cobalt (1938 a) and iron (1942), and which has been found satisfactory by other workers. This has greatly extended the range of powder pattern investigations, for suitable surfaces can be obtained without depending on the occurrence of smooth natural crystal surfaces of the desired orientation.

Attention here will be devoted mainly to the work of Elmore (1938 a) on cobalt, and of Williams, Bozorth and Shockley (1949) on iron, as the earlier work has often been reviewed, notably by Elmore in Bitter's *Introduction to Ferromagnetism* (1937, pp. 55–66). It is evident that, at the time of Elmore's writing, no satisfying interpretation had been given of any of the observed powder patterns in terms of a

magnetic domain structure proper, although the patterns were generally accepted as giving some kind of qualitative evidence for such a structure. Quantitatively relevant material was, however, available in the powder pattern photographs obtained by Kaya (1934 a, b) for iron, and by Kaya and Sekiya (1935) for nickel, as well as in earlier photographs by Bitter and others, though its significance was not appreciated till later. The occurrence on iron of three or four distinct types of pattern, as found by Kaya, was confirmed by Sixtus (1937 a), but only bare reference to this work can be made here, as also to the magnetic powder experiments of Snoek (1936) and Snoek and Louwerse (1937) on rolled nickel–iron. The first reasonably satisfactory quantitative interpretation of a powder pattern photograph was that by Elmore (1938 a), on the basis of the theoretical work of Landau and Lifshitz (1935), which, together with the comprehensive treatment of Néel (1944 b), provides the starting point for the detailed quantitative interpretation of particular observed patterns by Williams, Bozorth and Shockley (1949).

Formation of patterns. Although consideration has been given to certain special problems of the distribution and movement of colloid particles (e.g. by Elmore 1940 b), the elementary principles underlying the relation between a powder pattern and the magnetic structure at and near the surface seem usually to be regarded as so obvious as not to require discussion. Since powder patterns are easily misunderstood, however, brief comment may not be out of place. The force on the particles is proportional approximately (i.e. neglecting the effect of any permanent particle moment) to the gradient of H^2, and they therefore tend to move to regions in which H^2 has a local maximum. Since the particles are very small and very near the surface their motion is primarily controlled by the surface distribution of poles. The surface density of magnetic charge is equal to the magnetic intensity normal to the surface. Over regions in which the magnetization is tangential to the surface the force is zero; over regions in which there is a uniform normal component, there is a normal force attracting the particles to the surface, but no tangential force such as is required for pattern formation. Regions with different uniform surface pole densities may be differentiated to some extent by differences in the general density of the powder 'deposit'. The formation of a pattern line requires either a line of pole density differing in strength or sign from the surroundings on either side (e.g. the domain boundaries in the surfaces in Figures 3.3 and 3.4), or a boundary between two regions of differing surface pole density. The intensity of pattern lines, for a given colloid concentration, must depend on the saturation intensity of magnetization of the material (giving a measure of the attainable normal intensities) and on the effective widths of the transition regions (giving an inverse measure of the attainable local fields). From the comparative figures for domain boundary widths, b_0, in Table 3.2, clear patterns should be more readily obtainable on cobalt than on iron, for the second reason, and on either of these than on nickel, for both reasons. Materials with low anisotropy, and hence large domain boundary widths, would be expected to give poor patterns; in this connection it is noteworthy that on an unstressed crystal of permalloy (76 Ni, balance Fe) Williams, Bozorth and Shockley (1949, p. 171) could obtain no distinct patterns at all.

Patterns on cobalt. In his study of cobalt, Elmore (1938 a) used specimens containing a number of large crystal grains (of linear dimensions up to about 0·3 cm.). On these, surfaces were prepared approximating to prism planes and basal planes, parallel and perpendicular respectively to the hexagonal axis, the

easy axis. The specimen was mounted with the surface under observation horizontal on the pole of a vertical electromagnet, so that a field normal to the surface, outward or inward, could be applied. Some powder patterns obtained are shown in Figures 3.5 and 3.6 (Plate I). These patterns, which correspond to portions of those in the figures in Elmore's paper, are reproduced from original photographs kindly supplied by the author.

The three patterns (*a*), (*b*), (*c*) in Figure 3.5 correspond to the same region of the surface (approximately $0{\cdot}8 \times 0{\cdot}3$ mm.) with the applied field outward, zero and inward; and similarly the three patterns in Figure 3.6. Considering first the prism plane pattern in zero field, it is apparent that it is consistent with the theoretical scheme of layer magnetization illustrated in Figure 3.3, alternate layers being magnetized in opposite directions along the easy axis. The surface is slightly inclined to the easy axis, and there is a corresponding normal component of magnetization alternately outward and inward. An applied normal field has two effects: first, owing to the component along the easy axis, it gives rise to boundary displacement, with an increase in the volume (in cross section, of the width) of domains magnetized with a component in the field direction; secondly, the slight rotation of the magnetization vectors increases the surface pole density in these same domains, and decreases that on alternate domains. Thus, with an outward field, the width of domains in one of the two alternate sets increases, and the colloid density on them is relatively increased. With an inward field, the other set of domains is similarly affected. The colloid density remains greatest along the boundaries, but, starting from the zero field state, the colloid spreads in one direction or the opposite from any particular boundary according to whether the applied field is outward or inward. As is well shown in the photographs, in an applied field the colloid is practically cleared from alternate domains, and with reversal of the field, the clear and dark spaces interchange.

The beautiful patterns in the basal plane in Figure 3.6, in which the same interchange of light and dark regions is apparent, show that the domain structure in detail is much more complex than is suggested by the simple scheme of Figure 3.3. The transition from a prism plane to a basal plane pattern could to some extent be followed. On approaching a corner connecting these two planes, V-shaped line deposits appeared which multiplied the number of lines at the edge of the prism pattern; and on planes more inclined to the hexagonal axis than that of Figure 3.5, patterns were observed resembling those of Figure 3.6, but elongated in one direction. Elmore suggested that for a crystal magnetized to saturation along an easy direction, with a surface basal plane, reduction of magnetization might proceed by the starting of reverse regions from many centres on this basal plane, and their thread-like growth inwards, the threads fusing with threads from an opposite surface, and also laterally, eventually forming the layer-like structure over the greater part of the crystal. The theoretical work of Lifshitz (1944) shows that a decrease in energy may in fact be obtained by a branching of the layers near a basal plane much in the way that Elmore suggested, though quantitative calculations, particularly for multiple branching, would present formidable difficulties.

It was observed that the smaller grains gave the closer spacing in the prism plane patterns, in qualitative accordance with the theory. The spacing of the lines in Figure 3.5 is about 60 μ, a rough value of the relevant crystal length being 0·3 cm. This spacing is some three or four times greater than that calculated

from the simple formula (3.17), but the domain structure on the basal planes is so different from that assumed that this discrepancy, which is in the right direction, is not serious. Unfortunately, owing to the ill-defined crystal lengths, and the complexities of the closure domain structure, it is not possible, proceeding in the reverse way, to derive any really precise information about domain boundary energies from measurements of the patterns. None the less, through this work on cobalt, the relation between domain theory and observable powder patterns could at last justifiably be regarded as irrefutably established.

Patterns on iron. It is impossible to do justice in a short summary to the comprehensive work of Williams, Bozorth and Shockley (1949) and of Williams and Shockley (1949) on domain patterns on single crystals of iron. Fortunately, no more than an indication of some of the salient points is necessary, as these two papers should be read in their entirety by all who are interested in ferromagnetism, and also by others who are not yet. The material used consisted of carefully prepared single crystals of iron containing 3·8 weight per cent of silicon. Specimens were cut with surfaces approximately parallel to (100), (110) and (111) planes, and electrolytic polishing was carried out in a chromic acid–phosphoric acid bath. Patterns were obtained with a soap-protected colloidal suspension of magnetite. The patterns shown in subsequent figures are a small selection from a much larger number which, with the exception of Figure 3.12, have appeared in the papers just mentioned, or in an earlier preliminary paper by Williams (1947). They are, however, reproduced directly from original photographs kindly supplied for this purpose by the authors.

A number of methods were developed for determining the direction and sense of the surface magnetization in the areas between the powder pattern lines. A light scratch made on the surface results in the formation of poles if the magnetization is not parallel to the scratch. Figure 3.7 (*a*) shows part of a typical 'tree pattern' on a surface slightly inclined to a (100) plane. In (*b*), light parallel scratches have been made, with a ruling engine, in an EW direction; these are shown up by deposition of colloid only where the magnetization is NS. In (*c*), both NS and EW scratches have been made by a glass fibre brush. The interpretation of the pattern is given in (*d*). Even without scratches the directions of magnetization can be assessed from the form of colloid striations, which are always elongated in a direction at right angles to the field. This effect, shown in several of the figures, is particularly clear in Figure 3.11, where a more concentrated colloid has been used, and is probably due to a very slight unevenness in the electrolytically polished surface. (Over-etching gives the surface a microscopic appearance 'not unlike that of an orange peel'.) Confirmatory evidence is provided by subjecting the specimen to tension. Iron has a positive magnetostriction, and regions magnetized along the tension direction increase at the expense of those magnetized transversely, as shown in Figure 3.8. The sense of the magnetization could be determined by placing a long thin permanently magnetized probe on the surface under examination. The additional field due to the probe induces the formation of small additional reverse domains in a domain magnetized in the opposite direction, but has no effect on a domain magnetized in the same direction. The sense of the magnetization can ordinarily be found by observing the relative changes of area of different parts of the pattern in an applied field. Useful information can sometimes be obtained by the use of vertical as well as horizontal fields.

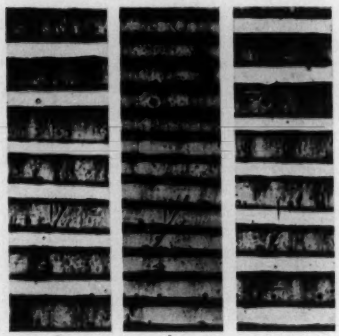


Figure 3.5. Patterns of magnetic colloid on cobalt surfaces approximately perpendicular to the basal plane.

a, *b*, *c*, applied normal field outward, zero, inward. (Elmore 1938)

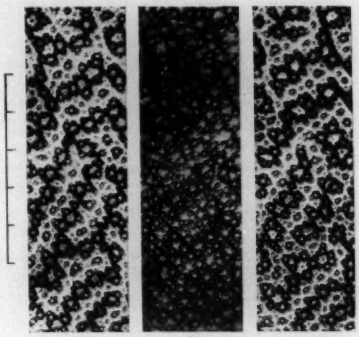


Figure 3.6. Patterns of magnetic colloid on cobalt surfaces approximately in the basal plane.

a, *b*, *c*, applied normal field outward, zero, inward. (Elmore 1938)

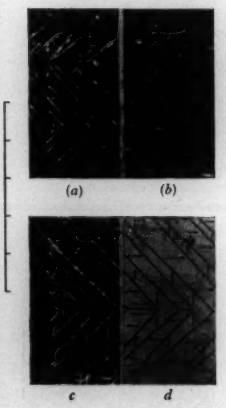


Figure 3.7. Patterns on (100) surface of iron.

Axes NS, EW. In *b*, parallel EW scratches have been made, in *c*, EW and NS scratches. The directions of magnetization corresponding to *c* are shown in *d*. (Williams 1947)

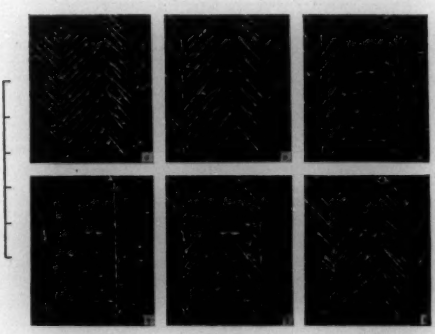


Figure 3.8. Effect of tension on patterns on (100) surface of iron.

Axes NS, EW. Tension, NS, increasing *a* to *d*, decreasing *d* to *f*, zero *a* and *f*. (Williams, Bozorth and Shockley 1949)

The scale divisions for all figures correspond to 0·1 mm.

With acknowledgments to the authors and to the *Physical Review*.

PLATE I.

To face page 120

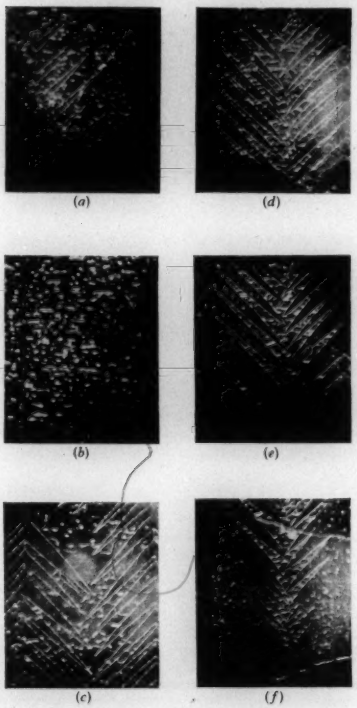


Figure 3.10. Patterns at successive positions, (*a*) to (*f*), on a curved surface.
Crystal axes NS, EW. (Williams, Bozorth and Shockley 1949)

Figure 3.13. Domain pattern on cold-rolled polycrystalline iron–silicon sheet. (Williams)

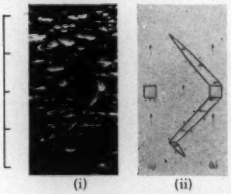


Figure 3.12. Domain structure round a hole in a crystal, (i) as observed, and (ii) interpreted. (Williams 1947)

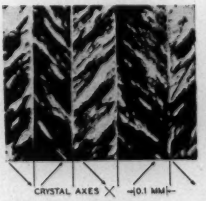


Figure 3.11. Pattern obtained with concentrated colloid on a (100) surface, magnetized in a [110] direction. (Williams, Bozorth and Shockley 1949)

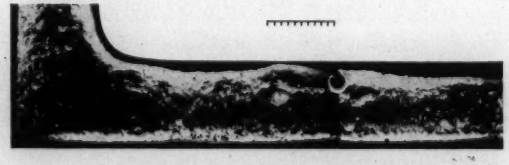


Figure 3.14. Composite of patterns on three adjacent areas of hollow rectangle single crystal. (Williams and Shockley 1949)

The scale divisions for all figures correspond to 0·1 mm.
With acknowledgments to the authors and to the *Physical Review*.

The patterns of Figure 3.8, reminiscent of spiky pinnate leaves, but usually referred to as 'tree patterns', are characteristic of (100) surfaces. The central line corresponds to a 180° boundary. Careful experiment showed that the branches occur only when the surface is slightly inclined to the (100) plane, and that they correspond to shallow superficial domains as represented in Figure 3.9, a slightly simplified version of a diagram given by Bozorth (1949). The twin branches (or leaflets) are partial closure domains (in the sense of §3 (i)), which reduce the magnetostatic energy associated with the surface poles by carrying flux through the crystal in a direction transverse to that of the magnetization of the main domains. This involves an increase in boundary energy (but not, in this case, in anisotropy energy). The minimum energy problem is extremely difficult, but an approximate treatment gives results in good agreement with the observations, such as those shown in the series of patterns in Figure 3.10, obtained at successive positions on a curved surface cut on a single crystal. In (*a*) the branches point upwards. This corresponds to Figure 3.9; the theory shows that the branches should point 'downhill' with reference to the (100) planes. In (*b*) the (100) plane

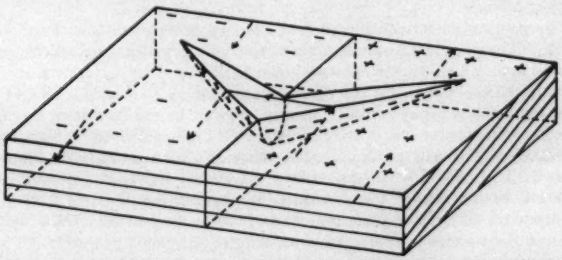


Figure 3.9. Interpretation of tree patterns. The lines on the side surface are traces of (100) planes.

is tangential to the surface, and only the centre lines appear, these corresponding to the 180° boundaries, with the magnetization on either side parallel to them. In (*c*) the branches point downwards, and as the slope increases the transversely magnetized region expands until in (*f*) only a relatively narrow jagged strip remains magnetized in the original direction. It is to be noted that in (*e*) and (*f*) the branched strip is what is left by the expansion of the original branches; it is a residual region on which there is surface magnetic charge, and the centre line does not, as in (*a*) and (*c*), correspond to a 180° boundary. A good indication of the directions of magnetization is given by the colloid striations.

A pattern obtained with concentrated colloid on a (100) surface magnetized in a [110] direction ($B = 15{,}900$) is shown in Figure 3.11. This corresponds to the top surface in Figure 3.4, and is clearly in agreement with Néel's theoretical treatment. On a (011) surface (the front surface of Figure 3.4) two types of domain surface were observed, whose relative areas changed with increasing field in the way predicted. The detailed character of the colloid deposition on (011) surfaces, and also on the edges of (100) surfaces shows, however, that the structure of the closure domains is much more complex than that suggested by Figure 3.4.

Many interesting patterns other than those illustrated here were obtained; studies were made of the effect of compression as well as tension; and the change

in spacing and in the form of the patterns with change of field were examined. Such matters as the course of spin orientation change in domain walls of the type illustrated in Figure 3.9 are fully discussed. In an appendix a theoretical interpretation is developed, in considerable detail, of the patterns. The observed forms are accounted for, and there is usually agreement between calculated and observed linear dimensions to a factor of about two.

There is, of course, scope for much further quantitative work, to which a first contribution has been made by Bates and Neale (1949), who have obtained excellent parallel line patterns on a (100) surface of a thin single crystal strip of silicon iron about 2 cm. wide and 4 cm. long, with the field as in Figure 3.4. The applied field was measured directly with a magnetic potentiometer (cf. Bates 1945). The spacing of the lines varied with the field at first rapidly and then more slowly, from about 300μ for $H=4$ to a fairly constant value of rather more than $100\,\mu$ for H ranging from 200 to 400. The agreement with Néel's theoretical curve is reasonably good, but further investigation is required before a precise comparison can be made. A fuller account of this work is in course of publication (Bates and Neale 1950).

The patterns so far considered have been those characteristic of the perfect crystal. Mention should also be made of the domain patterns round cavities and inclusions, of which an example is shown in Figure 3.12. It was first clearly shown by Néel (1944 d) that the magnetostatic energy associated with free poles on the surface of a cavity could be greatly reduced by the formation of suitable secondary domains round it. The pattern of Figure 3.12(i) conforms to the general type predicted by Néel. The directions of magnetization, shown in Figure 3.12(ii), which is self-explanatory, have been arrived at by careful study of the colloid striations. The formation of these secondary domain structures, here rendered visible, is of great importance in its bearing on the effect of inhomogeneities of structure or composition on ordinary magnetic properties.

Lest it should be thought that the effects of crystal grain boundaries are being entirely overlooked, a pattern obtained on an annealed cold-rolled polycrystalline iron–silicon sheet is shown in Figure 3.13. This photograph, hitherto unpublished, is due to Williams. The complex secondary domain structure suggests that the effect of grain boundaries on magnetic properties may be much more important than is sometimes supposed.

The picture frame crystal. This account of powder patterns may well be concluded by reference to the beautiful experiments of Williams and Shockley (1949) on a single crystal in the form of a hollow rectangle of the type already mentioned (§2 (ii)). This crystal, of 3·8% silicon–iron, with sides parallel to [100] type directions, had overall dimensions of $1{\cdot}9 \times 1{\cdot}3$ cm., a thickness of 0·074 cm., and a leg width of 0·102 cm. The surfaces were carefully ground parallel to (100) planes (to avoid the formation of superficial tree pattern domains) and polished for powder pattern observations. A pattern obtained after application of a field is shown in Figure 3.14. The salient features of this pattern are the diagonal line running across the corner, and the straight, slightly off-centre lines running down the limbs. The domain structure of this picture frame crystal is shown diagrammatically in Figure 3.15. After cooling from 1000° c., the crystal is fully magnetized and consists, apart from small superficial regions, essentially of four domains, one along each side, and separated by the 90° boundaries which show in the powder patterns as the diagonal lines at the corners. When a

suitably directed field is applied the 180° boundaries appear. Photographs taken while tracing a hysteresis loop showed that the position of one of these walls varied linearly with the resultant induction. In Figure 3.14 there is an inner wall due to excess material on the inside corner of the frame.

Although the motion of a 180° wall in a slowly changing field was relatively steady, the Barkhausen effect could be heard simultaneously with an amplifier and loudspeaker system. The movement is, in fact, slightly jerky. Superficial domain structures round small holes or other imperfections appear to unite with the wall and form irregularities in it. These secondary walls are extended as

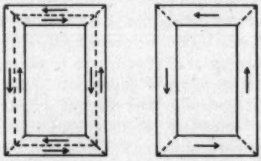


Figure 3.15. Domain structure in picture frame single crystal, with (100) surface plane, and sides parallel to [100] directions.
(*a*) After applying a field. (*b*) After cooling from 1,000° C.
(*After* Williams and Shockley 1949.)

the main wall progresses, until they suddenly break. A new structure is formed round the imperfection, and the main wall straightens out. Photographs are given in the original paper illustrating this process.

A method is incidentally proposed by which it is hoped to make an almost direct determination of the domain boundary energy, in effect by measuring the curvature of the domain boundary in suitably disposed magnetic fields. Apart from this, these experiments have shown, in an ideally simple case, that there is a direct correlation of the domain structure with the magnetization, and they have provided immediate visual information about primary and secondary processes taking place when a ferromagnetic is magnetized.

3 (iii). *Discontinuities in Magnetization: the Barkhausen Effect*

General survey. When the field applied to a ferromagnetic is changed continuously, in general the change in magnetization is, at least in part, discontinuous. It is convenient to distinguish the small scale discontinuities in ordinary materials, first observed by Barkhausen (1919) by the frequently repeated method involving the use of a search coil, amplifier and loudspeaker, from the large scale discontinuities which occur in materials under stress (§ 3 (iv)), and which appear to have been first investigated systematically by Forrer (1926).

During the ten to fifteen years following its discovery, numerous investigations were made on the Barkhausen effect, culminating in the work of Dillinger and Bozorth (1929–32) (I, pp. 61–2). This work was far more thorough than any that had preceded it, but greater finality has been attributed to the tentative conclusions as to domain sizes and the aggregate extent of the discontinuous changes than was, perhaps, justified, with the result that, until this type of investigation was taken up again by Bush and Tebble (1948) little further systematic experimental work was carried out on the central problem. Brief mention should, however, be made

of a few of the investigations in the intervening period. Following earlier work by McKeehan, a method for determining the directions of the discontinuous changes of magnetization was developed by Clash and Beck (1935). A rotating field was applied to a single crystal (100) disc of silicon–iron, and the voltage impulses due to the discontinuities were picked up by two coils at right angles, amplified, and fed to an oscillograph in such a way that traces were obtained corresponding in direction to the original discontinuous changes in magnetization. It was found that, in medium applied fields, practically all the large changes could be attributed to simple reversals along one of the axes of easy magnetization. The distribution in detail, however, is complicated and not readily explained. Extensive oscillographic studies of the Barkhausen effect have been made by Murakawa (1936, 1937), who noted the occurrence, as the field changed gradually, of concentrated groups of peaks, which, it was suggested, might arise from a series of changes occurring in rapid succession along a 'filament' in the wire specimen. Takagi (1937) examined the effect at different temperatures for nickel and a ferronickel. He found that the magnitude of the discontinuous changes diminished with increasing temperature, and that at the same time prominent groups of peaks, which could be 'recognized', were displaced to smaller fields. Heaps (1941) concluded, from an investigation of the sound waves due to the magnetostrictive changes for jumps through angles other than 180°, that in nickel the 180° changes were predominant for small values of I/I_0, but that changes through other angles occurred over most of the range. In an investigation of the low field region for iron, nickel and a ferronickel, Montalenti (1948) found that the Rayleigh law ceases to hold at approximately the value of the field for which the Barkhausen effect first becomes noticeable.

The Bozorth-Dillinger work. Although a description of the work of Bozorth and Dillinger is hardly necessary here, as it has been so often summarized, certain aspects of it call for brief comment. In view of the obvious difficulties of more direct measurements of the sizes of discontinuities sufficiently extended in range and number to be representative, the indirect method used had the great advantage that, in principle at least, it enabled the mean size of the discontinuities, and their contribution to the change of magnetization, to be determined for any part of the magnetization curve. Apart from the admitted approximations involved, however, the significance of 'mean size' or the associated 'mean volume', $\bar{v}$, is somewhat problematical. The mean size must depend on the smallest discontinuity which can be detected, and the negative statement that discontinuities corresponding to volumes, in iron, of 10^{-13} cm^3 or less would not appreciably affect the measuring system does not clarify the position, for the effect of such discontinuities would almost certainly be far below the inevitable noise level of the amplifier. It is thus by no means clear just what information is conveyed by the well known curves (e.g. I, Figure 2.7, p. 62) giving the variation of $\bar{v}$ over the magnetization curve, and showing a maximum in the region of the coercive field. From the results generally it was concluded that 'the changes in magnetization which take place suddenly in large groups of atoms account quantitatively for the whole change in magnetization which corresponds to the steeper part of the hysteresis loop'. This is, indeed, rendered plausible by the curves shown for some materials (e.g. hard drawn permalloy) but for others the actual observed values for the aggregate discontinuous change (i.e. without extrapolating to zero rate of field change) differ from the observed total change by a factor of ten or more.

Although information about the character of the Barkhausen effect was so greatly extended by the work of Bozorth and Dillinger, it is clear that wide scope remained for further investigations.

The counting method. The work of Bush and Tebble (1948) may be regarded as an extension and refinement of that of Bozorth and Dillinger, made possible by the methods of counting electrical pulses developed in the intervening period. The basic idea was, using a similar field changing device, to pass the amplified pulses from the search coil through a discriminator and scaling unit to a recorder, so that the size–number distribution of the discontinuities could be obtained. Apparatus for this purpose was set up, and some preliminary results obtained with it are given in the paper mentioned. In the light of the experience gained, the apparatus was rebuilt with many improvements in detail, as described by Tebble, Skidmore and Corner (1950), who give results for hard drawn iron, a specially treated iron containing large crystals, and annealed nickel.

Pulse shapes and design of apparatus. A necessary part of the investigation was a theoretical and experimental study of the pulse shapes due to the discontinuities. The full analysis is complicated, but in broad outline the pulse shape from the amplifier (provided certain conditions as to its frequency range are satisfied) due to a single Barkhausen discontinuity depends on the characteristic times associated with the discontinuous process itself, τ_B, with the decay of eddy currents in the specimen, τ_E, and with the search coil and amplifier input circuit, τ_C. The first, τ_B, the time taken for the initial spontaneous boundary movement, may be estimated as usually being of the order 0·1μsec. The second, τ_E, is proportional to $\mu\sigma a^2$, where μ is the reversible permeability, σ the specific conductivity and a the radius of the specimen. The detailed form of the pulse, and its amplitude, considering only the τ_B and τ_E effects, depend on b/a, where b is the distance of the discontinuity from the axis. The third, τ_C, is in principle calculable from the L, C, R characteristics of the coil circuit. The final form of the pulse is mainly controlled by the longest of these three characteristic times, and in general τ_E and τ_C are both much greater than τ_B. For obtaining oscillograph traces which correspond to pulse shapes in the specimen, the necessary condition is $\tau_C \ll \tau_E$. This was readily fulfilled by suitable coil design (with a particular coil used, the value of τ_C was about 6μsec.), and by using specimens of slightly larger radius than for the counting experiments (e.g. with a hard drawn iron wire, for which a radius 0·08 cm. was suitable, τ_E was about 177μsec.); the distribution of pulse widths was found to be in good accord with the detailed theory, bearing in mind the uncertainty as to the effective value of the reversible permeability, which, under the conditions of rapid change, must be appreciably smaller than the measured values at low frequencies. In the counting experiments, in which use is made of a voltage amplitude discriminator, it is obviously necessary that the pulse amplitude should be proportional to total flux change (i.e. to the change in magnetic moment) associated with the Barkhausen jump; this can be ensured only if the opposite condition, $\tau_C \gg \tau_E$, is satisfied. Specimens of smaller radius were used (the specimens, about 7·5 cm. long, usually had a radius of about 0·02 cm. and the τ_E values for the three materials, Fe1, Fe2 and Ni, were about 10, 46 and 17μsec.) and a larger search coil, some 10 cm. long and with 50,000 turns, giving a τ_C value of about 200μsec.

In the general design of the complete system, careful attention was given to such matters as signal to noise ratio, and the use of a sufficiently slow rate of change

of field (usually in the range 1 to 20×10^{-4} oe.sec^{-1}) to minimize coincidences; the original paper must be consulted for details of the various component parts of the apparatus. It must be sufficient to state that the amplifier, with an adjustable frequency range, had a maximum gain of about 2×10^5; that the pulses selected by the discriminator were converted to a uniform size and passed via a 1,024 scaling unit (i.e. 10 successive scale-of-two stages) to a telephone recorder; and that, after calibration with an artificial pulse unit, counts were made of pulses corresponding to changes of magnetic moment, ΔM, ranging from about $0{\cdot}3$ to 40×10^{-6}, the lowest discriminator setting being well above the noise level of the apparatus. The complete examination of a single specimen, it may be mentioned, requires a considerable time, for, with the necessarily slow rate of change of field, a single run over a half cycle of a magnetization curve for one discriminator setting may occupy up to twenty-four hours.

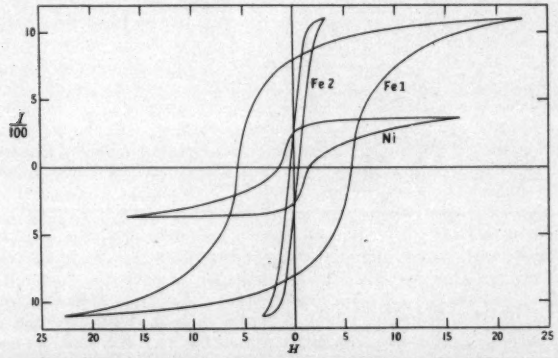


Figure 3.16. Representative hysteresis curves of materials used in Barkhausen effect investigations. H, applied field. Fe1, hard drawn iron; Fe2, annealed iron, large grained; Ni, annealed nickel.

(*After* Tebble, Skidmore and Corner 1950.)

Experimental results. The primary experimental results may be expressed in the form of a curve, the (number, level) curve, giving the number of discontinuities, N, per unit volume, for which the change in resolved magnetic moment is greater than ΔM, obtainable from the discriminator setting, over a range of the magnetization curve corresponding to the total change of field. From this curve the contribution of the discontinuous changes, over any ΔM range included, to the total change in magnetization is at once obtainable by integration. Differentiation of the curve yields a distribution curve, giving $dN/d(\Delta M)$ against ΔM. Finally, multiplying the ordinates of this curve by ΔM, a contribution curve is obtained, $\Delta M\{dN/d(\Delta M)\}$ against ΔM, giving the contribution, per unit range of ΔM, of the discontinuous to the total change in the magnetization. Hysteresis and contribution curves for the three materials investigated in most detail are shown in Figures 3.16 and 3.17. The iron content of the hard drawn iron (Fe1) was better than 99·9%, the main impurities being C (0·03) and Mn (0·05); the annealed iron (Fe2) was prepared from it by decarburization followed by the

standard stretching and annealing process to produce large crystals; the nickel, chosen because related magneto-thermal investigations had been carried out on similar material, had a purity of about 99·6%. For convenience in graphical representation, the abscissa for the contribution curves is taken as $(\Delta M/2I_0)$, so the horizontal scale gives directly the minimum volumes associated with the discontinuities. The lower limit of the curves is fixed by the minimum value of ΔM for which reliable counts can be made ($\Delta M \simeq 0{\cdot}3 \times 10^{-6}$), corresponding to a minimum volume of slightly less than 10^{-10} cm³ for iron, and a value about three times as great for nickel.

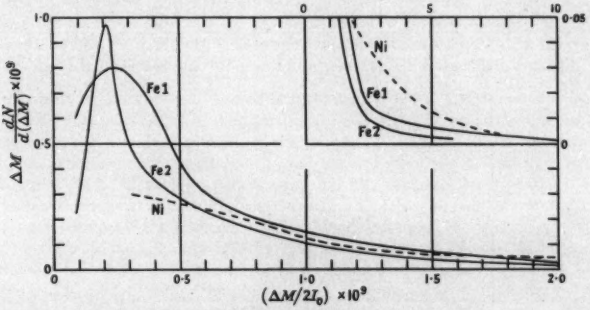


Figure 3.17. Contribution curves for iron and nickel, showing the contribution to the total change of magnetization from discontinuous changes of moment, ΔM.
(*After* Tebble, Skidmore and Corner 1950.)
$(\Delta M/2I_0)$ gives the minimum volume associated with the change in resolved moment, ΔM.
Values adopted for $2I_0$: Fe, 3,440; Ni, 970.

Table 3.3. Contribution of Barkhausen Discontinuities to Change in Magnetization. (Tebble, Skidmore and Corner 1950.) (See text for symbols and details of materials.)

Material	Range of field	ΔI_T	Range of $\Delta M \times 10^6$	ΔI_B	ΔI_{rev}	$\Delta I_B/\Delta I_T$
Fe1	±18·6	2150	0·3–35	1844	180	0·86
Fe2	± 3·0	2180	0·3–20	1250	110	0·57
Ni	±14·9	700	0·3–12	326	115	0·47

The contribution from the Barkhausen discontinuities, ΔI_B, actually counted over the range investigated to the total change in magnetization, ΔI_T, is shown in Table 3.3, and also the reversible contribution, ΔI_{rev}, as estimated from measurements of the reversible susceptibility (Tebble and Corner 1950). Some representative curves showing the measured discontinuous and continuous contributions to the change of magnetization with field for nickel are shown in Figure 3.18. For hard drawn iron, from the numbers in Table 3.3, the reversible contribution accounts for 0·08 of the change in magnetization, and the measured discontinuities for 0·86. Slight extrapolation (see Figure 3·17) to cover the smaller discontinuities, gives a further 0·04. Although the closeness of the overall agreement in this case must be regarded as to some extent fortuitous, it none

the less gives convincing evidence that considerable reliance can be placed on the substantial accuracy of the size–number determinations. For nickel, extrapolation would be more uncertain, but an overall discrepancy of perhaps 20% is indicated. For annealed iron (Fe2) the discrepancy is greater, about 35%, after extrapolation. There is a possibility that the reversible contribution is underestimated, but it is unlikely that the error is large. The most probable source of the discrepancy is in the overlapping of the pulses due to temporally close discontinuities, an effect likely to be more persistent, even with the slowest practicable rates of change of field, the higher the permeability of the material. It was observed, moreover, particularly in annealed iron, that close groups or bursts of pulses occurred, and that although the intervals between the groups increased with decrease in the rate of change of field, the apparent interval between the distinguishable pulses in these groups was not appreciably affected. This

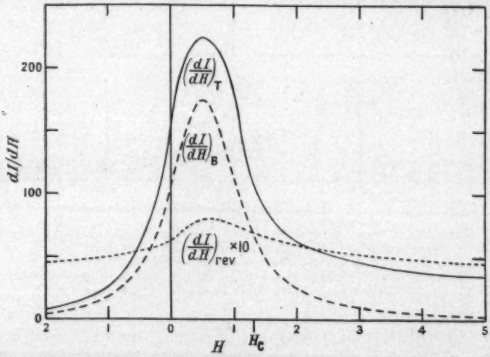


Figure 3.18. Change of magnetization with field, $(dI/dH)_T$, for nickel, and measured contributions from Barkhausen and reversible processes, $(dI/dH)_B$ and $(dI/dH)_{rev}$. (*After* Tebble, Skidmore and Corner 1950.)

apparent effect of one discontinuous process on others, combined with incompleteness of resolution of close pulses, is probably sufficient to account for the somewhat lower counts for specimens of higher eddy current time constants, τ_E. The check on the experimental work made by bringing together curves such as those of Figure 3.18 is very severe, and stress should be placed primarily on the general agreement rather than on discrepancies in detail.

From the results on which Figure 3.17 and Table 3.3 are based, it is clear that the bulk of the discontinuous change of magnetization occurs, in iron, by changes of moment, ΔM, ranging for 0·3 to 30×10^{-6}, corresponding to minimum volumes from about 10^{-10} to 10^{-8} cm^3, and numbers of atoms from about 10^{13} to 10^{15}, and that in nickel (from Figure 3.17) the volumes are much the same, and the changes in moment some three times smaller. The size distribution of the discontinuities was examined, not only for the range covered as a whole, but also for successive parts of the range. No definite indication of any significant variation in the distribution was found. Thus, although the contribution from the discontinuities

rises to a maximum roughly in the region where $(dI/dH)_T$ is a maximum (Figure 3.18 is representative in this respect) the 'average volume' of the discontinuities does not, in sharp contradiction to the findings by the indirect method (I, Figure 2.7, p. 62). It should be re-emphasized that the concept of 'average volume' applied to Barkhausen discontinuities requires very careful definition if it is to be useful. If the contribution curves are of the type shown for Fe1 and Fe2 in Figure 3.17, the volume corresponding to the peak contribution (about 2·5 and 2×10^{-10} cm^3 in these two cases) might be taken as characteristic; alternatively, a volume, $v_{\frac{1}{2}}$, such that half the total discontinuous charge came from smaller, and half from larger volumes, might be given as an appropriate characterizing magnitude. Characteristic volumes so defined appear to be some ten times smaller than the maximum average volumes obtained by the indirect method; average volumes in the usual sense would, of course, be very much smaller still.

Significance of sizes of discontinuities. The experimental results for the size distribution of Barkhausen discontinuities having been considered at some length, it must now be admitted that the precise significance of these sizes (or, more specifically, of the volumes of the groups of atoms in which the changes occur) is at present very uncertain. It was originally thought that the Barkhausen volumes could be identified with domains, but this is certainly not generally true if the term 'domain' is used in the sense here adopted. More non-committally, and indeed almost incontrovertibly, a Barkhausen volume has for some time been considered as a volume traversed by a domain boundary (or part of a domain boundary) in moving spontaneously from one equilibrium position, which becomes unstable at a critical value of the effective field, to a new stable equilibrium position (as represented diagrammatically in Figure 4.1 below). The range of these movements must, it would seem, be limited by static irregularities in the crystal structure, giving conditions for localized boundary movement essentially distinct from those determinative of the general relatively large-scale movement of the primary domain boundaries. It is a tempting speculation that Barkhausen volumes may be closely related to the slightly disorientated mosaic blocks in metallic crystals, the volumes of which appear to be of much the same order of magnitude (e.g. Barrett 1943, pp. 219–23; Lacombe and Berghezan 1949). An extension of the work of Williams and Shockley (see § 3 (ii)), in which the changing powder pattern and the Barkhausen effect are observed simultaneously, offers great promise, though it should be mentioned that the reorganization of the secondary domain structure round an imperfection shown in one of their photographs appears to involve a volume very much greater than those for typical Barkhausen discontinuities. The whole problem is a fascinating one calling for much further investigation, the bearings of which may be by no means limited to ferromagnetic materials. In recent years a much clearer understanding of the Barkhausen effect has been gained, and it seems likely that more complete and satisfying answers to some of the outstanding questions will not be long delayed.

3 (iv). *Related Investigations*

Large discontinuities. Among other types of investigation in the general field of domain structure there are two of particular interest and importance, to which some reference must be made. The first of these is the study of large discontinuities in magnetization, such as may be obtained in wires of materials with positive magnetostriction under tension. With increasing tension, the wire axis

approximates more and more closely to an easy axis, the free energy associated with the strain anisotropy being given by

$$F_S = \tfrac{3}{2}\lambda Z \sin^2\phi, \qquad \ldots\ldots(3.19)$$

where λ is the saturation magnetostriction coefficient, Z the tension, and ϕ the angle between I_0 and the axis. The relation holds more closely the greater the ratio $\lambda Z/K$, where K is the magneto-crystalline anisotropy coefficient. The hysteresis loops become approximately rectangular, and reversal of magnetization may occur in a single 'Barkhausen' discontinuity, the change of magnetization being $2I_0$. The work of Sixtus and Tonks (1931–33), Preisach (1932) and others (already briefly reviewed in I, pp. 62–4) shows that the initiation of a reversed region in a wire initially saturated in one direction requires a starting field, H_s, usually greater than the critical field, H_0, required to maintain the propagation of the reversed magnetization along the wire. Measurements were made of H_s and H_0 for different stresses, and of the propagation velocity under various conditions, the materials examined usually being ferronickels in the composition range 5 to 80% nickel. The effect of combined tension and torsion was examined by Reinhart (1934 a, b). It was later found by Sixtus (1935) that the extent of growth of a reversed region could be controlled by limiting the time of application of H_s, even though the steady field was maintained at a value somewhat greater than H_0. Thus, 'frozen-in' reverse domains may be obtained, their dimensions at least approximately determined, and the conditions for their growth examined. The measurements justify their being treated theoretically, with good approximation, as ellipsoids of rotation. The manner of growth of a reversed domain, laterally and longitudinally, as dependent on the applied field, can be estimated by considering the change in the external field energy, the boundary energy and the demagnetizing field energy. The extensive experimental measurements show satisfactory agreement with the detailed theoretical calculations (Döring 1938, Döring and Haake 1938, Haake 1939). This work is very fully reviewed in Becker and Döring's book (1939, pp. 182–7, 192–203). The results of these studies of the growth process for large domains are not only of interest in themselves, but they also have some bearing on the more general question of boundary movements in ordinary materials. The difference in the large and the small discontinuities is, however, not merely one of scale. The conditions of high and uniform stress under which the large discontinuities occur are very specialized, and the shapes of the large reverse domains are very different from those of domains in ordinary materials. The study of large discontinuities does not therefore solve the problem of small discontinuities, though it does bring into clear focus some of the factors which must be taken into consideration. It is perhaps relevant to mention that in an investigation of the approximately rectangular hysteresis loops of nickel under compression, it was found by Heaps (1936) that as the magnetization changed over the nearly vertical steep part of the curve, the ordinary Barkhausen effect was apparently much the same as with unstressed material.

Small particles. The second group of investigations is on the behaviour of ferromagnetic bodies (or assemblies of such bodies) at least one of whose dimensions is very small, that is, with thin films, 'needles', or small particles. For bodies of ordinary size, the magnetostatic energy associated with a state of uniform magnetization is usually reduced by the formation of domains; but below a

certain size the increase of energy due to boundary formation, depending on surface, will exceed the reduction of magnetostatic energy, depending on volume, and no boundaries will be formed. By determining the minimum energies associated with various physically plausible domain boundary arrangements (using arguments similar to those outlined in § 3 (i)) and comparing these with the energy when the body is uniformly magnetized, Kittel (1946) has made estimates of the critical sizes below which the single domain state is energetically favoured. Using, for simplicity, a standard value of 3 erg cm^{-2} for the boundary energy, and the magnetization and anisotropy values for iron, he obtains the following critical dimensions: for film thickness, 3×10^{-5} cm.; for the edge of a cubical particle, $1{\cdot}5 \times 10^{-6}$ cm.; and for a needle, values of the diameter varying with the ratio, m, of length to diameter, from $2{\cdot}8 \times 10^{-6}$ cm. for $m = 5$ to $2{\cdot}4 \times 10^{-5}$ cm. for $m = 20$. Somewhat similar calculations have been made by Stoner and Wohlfarth (1948) for particles in the form of ellipsoids of rotation. Using the values of $\alpha_1 a_1^2$ in Table 3.2, it is found that spherical particles are almost certainly single domain for diameters (in 10^{-6} cm.) below 2·4, 5·2 and 3·2 for iron, nickel and cobalt respectively; for prolate spheroids the corresponding lengths of the minor axis increase with the dimensional ratio inversely as the square root of the demagnetization coefficient.

Perhaps the most direct experimental evidence for the single domain character of small particles is provided by the magnetization curves for colloidal iron oxides (Elmore 1938 c). The curves are of the Langevin type, showing an approach to saturation in fields of a few hundred oersted, and for the magnetite sol are consistent with particles of mean diameter $3{\cdot}4 \times 10^{-6}$ cm., and with a 'permanent moment' corresponding to about 44% of the bulk magnetization of magnetite. In the older work on thin films and small particles, summarized by Kittel, much higher values for the coercivity and lower values for the initial permeability are recorded than for the bulk material. From much of the earlier work, however, for various reasons, it is difficult to draw any very definite quantitative conclusions. More recently Drigo and Pizzo (1948) have given a preliminary account of measurements on thin films of the ferromagnetic elements electrolytically deposited on wires or discs of copper. They find that the coercivity increases with decreasing thickness, and that Barkhausen discontinuities disappear for thicknesses (in 10^{-6} cm.) below 13, 8 and 12 for iron, nickel and cobalt respectively. An important application of the general ideas on single domain particles is to the interpretation of the properties of the high coercivity alloys and powder materials used for permanent magnets, to which further reference is made later (§ 4 (iv), § 5 (iv)).

Finally, mention may be made of some interesting work by König (1948) on thin films, though it is only indirectly relevant to the theme of the present section. Films of iron, some 25×10^{-8} cm. in thickness, were deposited on collodion at a low temperature. The degree of ferromagnetism of the film could be estimated from the Faraday effect in a high magnetic field, and crystal sizes from the line widths in electron diffraction patterns. Crystal formation occurs on warming the film from a low temperature. Magneto-rotation could first be detected when the estimated linear dimensions of the crystals were about 11×10^{-8} cm., corresponding to about four lattice spacings. This indicates that the minimum crystal size necessary for the development of ferromagnetic characteristics corresponds to

some 64 unit cells, or a volume of about $1{\cdot}5 \times 10^{-21}$ cm^3, containing some 128 atoms (cf. Becker 1948). From the evidence considered in this section ferromagnetic domains are ordinarily very much larger than this, and, as has been tacitly assumed throughout, there is no reason to suppose that there is any appreciable variation of the degree of spontaneous magnetization over the relevant range of domain size.

§ 4. THE ANALYSIS OF MAGNETIZATION CURVES

4 (i). *Introductory*

The two preceding long sections on single crystals and domain structure have been concerned with the fundamental processes of change of magnetization which underlie the forms of magnetization curves generally. In view of the fact that reasonably reliable values are available for magneto-crystalline anisotropy coefficients (§ 2 (ii)), and that there is remarkably close agreement between theoretically calculated and observed magnetization curves for single crystals (§ 2 (iii)), it might seem that an analysis (or synthesis) of the rotational contribution to the change of magnetization in polycrystalline material would be fairly straightforward. Even if the grains were individually ideal crystals, however, it is very difficult to make proper allowance, in an averaging process, for their mutual magnetic interaction, and, indeed, until very recently this problem has been almost entirely ignored. Conversely, given the applied field and the bulk magnetization of the specimen, only the roughest estimate could be made of the effective field for particular grains, the shapes of which must in general give rise to a distribution of magnetization physically as complex as it is mathematically unmanageable. Apart from uncertainties arising in this way, which should be borne in mind, though they may not always be serious, the grains, in general, are by no means ideal crystals. They have the same kind of imperfections in the form of localized heterogeneities, depending on composition, and thermal and mechanical treatment of the material, as have single crystals, but in addition are subject to a much greater degree of strain or deformation, so that the effective anisotropy of a grain may differ enormously from that of the ideal crystal. Much the same considerations apply in connection with boundary movement processes.

It follows at once that a knowledge of the effects of stress on magnetization, or conversely of the magnetostrictive changes accompanying magnetization, are of great importance in connection with the general interpretation of magnetization curves. Many of the basic ideas in this field, and much of the related experimental work, date from the early thirties (cf. I, pp. 61–65), and need not be considered in detail in this Report (see, however, § 3 (iv)). Such consideration is the less necessary, in that the subject has been dealt with, in varying degrees of detail, in many review articles and books, and particularly in Becker and Döring's *Ferromagnetismus* (1939). (This book will be referred to in this section as BD.) In fact, since about 1931 until very recently, the outlook on magnetization curves has been dominated by the 'strain theory'. The general formal theory of the joint effect of stress and field is very complicated (BD, pp. 132–146), but partly owing to this complication, and partly owing to the inadequacy of experimental data on the various elastic and magneto-elastic crystal coefficients, use is made in applications almost exclusively of a few simple approximate relations (BD, p. 145). The most important of these is the relation for the variable part of the

magneto-elastic free energy, F_s, of a ferromagnetic with isotropic saturation magnetostriction, λ, subjected to a tension, Z, namely

$$F_s = \tfrac{3}{2}\lambda Z \sin^2\psi, \qquad \ldots\ldots(4.1)$$

where ψ is the angle between the magnetization vector, I_0, and the direction of tension. The success of the application of (4.1) and similar relations in the coordination of the behaviour of ferromagnetics under applied stress is unimpeachable; the apparent success in the coordination of the low field properties of diverse types of ferromagnetics in terms of hypothetical distributions of internal stress is even more remarkable, but in the light of recent criticisms (see below) it must be regarded as at least in part illusory. There is always a danger, as pointed out by Brown (1938 b), 'of misinterpreting an agreement between theory and experiment as evidence for the particular forces or mechanism assumed in the theory'. In a series of papers on reversible susceptibility and related effects, Brown (1937–39) has developed with considerable success a statistical theory from which useful relations are obtained which are independent of the precise mechanism by which the equilibrium state of magnetization is established. It must be remembered, however, that some of the most characteristic features of the behaviour of ferromagnetics in low and moderate fields are due to the fact that the normal states are in general not equilibrium states of lowest free energy for the system as a whole. The problem is one of microscopic metastability, to which the methods of statistical mechanics are hardly applicable.

It is peculiarly difficult at the present time to give a coherent account of work on what may be called the 'analysis' of the ordinary magnetization curve, using that term in a broad sense to imply the determination of the character of the elementary changes occurring over any part of the various types of magnetization curves (e.g. initial, normal or 'commutation' cyclic curves) for a particular material, coupled with an explanation of the very diverse magnitudes of characteristic quantities or coefficients (e.g. remanence, coercivity, initial and maximum susceptibility or permeability) for different materials. The difficulty is largely due to the new knowledge recently gained, particularly on domain structure. This has shown that most of the earlier treatments of the magnetization curve problem are defective, as being based, at least in part, on physical assumptions now known to be untenable. Little useful purpose would be served by describing in detail theoretical treatments which the authors themselves would not now regard as satisfactory. This does not mean that the earlier treatments are valueless, for large parts of them will be incorporated in new treatments. For the most part, however, these new treatments have yet to be developed. The presentation of earlier work in this section will, therefore, be relatively brief and selective; the main aim will be to indicate how certain aspects of the problem of analysis of curves have been, or may be, approached in the light of the new knowledge.

From a theoretical standpoint, a distinguishing feature of the newer outlook is the recognition of the importance of purely magnetic interactions, previously largely overlooked. This is apparent in the treatment both of single crystal magnetization curves (§ 2 (iii)), where an important factor is the demagnetizing field of the crystal as a whole, and of domain structure (§ 3(i)), where the small scale patterns are largely conditioned by a minimization of the energy due to internal free poles. It will be apparent again in the supplementary topics discussed below in connection with boundary movements (§ 4(ii)) and rotations (§ 4 (iii), (iv)).

In §4(v), recent work on magneto-thermal effects is briefly reviewed as it is believed that experimental and theoretical studies of the changes of temperature accompanying magnetization provide a most valuable means of obtaining information about the character of the magnetic changes occurring additional to that obtainable from the magnetization curves themselves. In this section little reference to experimental methods is made, as the methods are for the most part standard and well known; nor is it considered necessary to give a formal description of the various types of magnetization curve.

4(ii). *Boundary Movements*

Reversible and irreversible boundary movements. Most of the factors which have been suggested as determinative of the extent and character of boundary movements can conveniently be indicated with reference to Figure 4.1. This particular figure is taken from a paper by Stoner and Rhodes (1949), but similar diagrams have often been given.

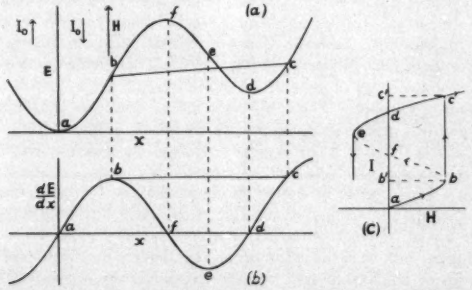


Figure 4.1. Diagrams illustrating reversible and irreversible boundary movements. The boundary is parallel to the yz plane, with its central plane passing through the energy minimum at a in zero field. As the field increases the boundary moves in the direction of increasing x. Corresponding points in (a), (b) and (c) are marked by the same letters.

(Reproduced, by kind permission, from the *Philosophical Magazine*.)

For simplicity a 180° boundary is considered. The curve in Figure 4.1 (a) represents a possible form for the dependence of the internal free energy of the material on the position of this one boundary. Using the symbol E for the energy per unit area of the boundary (but without necessarily implying that the energy is localized in the boundary itself), and assuming that in zero field the boundary is in equilibrium at $x=0$, the condition for equilibrium in a field H is

$$\frac{d}{dx}(E-2HI_0x)=0, \qquad \ldots\ldots(4.2)$$

or

$$2HI_0=dE/dx. \qquad \ldots\ldots(4.3)$$

As the field increases the boundary moves reversibly to the right until dE/dx reaches a maximum at b when the equilibrium becomes unstable. Once the corresponding value of the field, namely

$$H_0=\frac{1}{2I_0}\left(\frac{dE}{dx}\right)_{\max}, \qquad \ldots\ldots(4.4)$$

is passed, the boundary moves spontaneously (i.e. without further increase of field) to a new position of equilibrium at c, the change from b to c corresponding to a Barkhausen discontinuity. With further increase of field reversible movement is resumed. The course of the changes when the field is decreased will be clear from the figure. The diagram (c) shows how the boundary movement contributes to the reversible and irreversible susceptibility (ranges *ab* and *bc* respectively) in increasing field, and to the reversible susceptibility again (range *cde*) in decreasing field. The manner in which remanence, coercivity and hysteresis can arise from boundary movement are all illustrated. The assumption of a not unduly specialized type of dependence of energy on boundary position thus leads to a formal coordination of a wide variety of effects associated with a single elementary process, and it is not unreasonable to suppose that the bulk characteristics of ferromagnetics might arise from the superposition of a large number of similar processes. A variation of internal energy with boundary position must arise, in the last resort, from deviations from uniformity of the material, corresponding to a variation in internal stress or composition, and the theoretical problem is that of linking the two together. There have been three distinct lines of approach (within the same formal scheme which has been outlined), not necessarily mutually incompatible, and the treatments, as developed, may be referred to, for brevity, as the 'strain', 'inclusion' and 'disperse field' theories.

The strain theory. The basic idea of the strain theory, as first put forward in a systematic way in the present connection by Kondorski (1937) is that the variation of energy with boundary position is mainly due to the direct effect of the spatial variation of local stresses on the energy of the boundary itself. Thus, if the boundary energy, in the absence of stress, is given by

$$\gamma = \gamma_K = gK^{\frac{1}{2}}, \qquad \ldots\ldots (4.5)$$

where K is a magneto-crystalline anisotropy coefficient, and g a coefficient involving the exchange energy (see § 3 (i) for details), with internal stress Z (having the character of a tension or compression) the energy becomes

$$\gamma = \gamma_Z = g(K + 3b\lambda Z/2)^{\frac{1}{2}}, \qquad \ldots\ldots (4.6)$$

where λ is the saturation magnetostriction, assumed isotropic (cf. (4.1)), and b is a numerical factor depending on the orientation of Z relative to the boundary. In equations (4.2) to (4.4), E is replaced by γ. The critical field, H_0, is proportional to the maximum stress gradient, and, as may readily be shown from the relations given in § 3 (i), to the effective width, δ, of the boundary. In Kondorski's treatment it is assumed implicitly that the variation of Z over the width of the boundary is small. This limitation is removed in a development of the treatment by Kersten (1938 a, b), who considers a stress inhomogeneity extending over a length l (perpendicular to the plane of the boundary), in which the stress rises from zero (or from the mean value over the whole of the relevant region) to a maximum ΔZ. It may then be shown (cf. Kersten, 1938 a, b, or BD, p. 207) that the critical field is given by

$$H_0 = p_0(\lambda \Delta Z / I_0), \qquad \ldots\ldots (4.7)$$

where p_0 is a numerical factor proportional to l/δ for $l/\delta \ll 1$ and to δ/l for $l/\delta \gg 1$, and having a maximum value, of order unity, for $l \simeq \delta$. For a stress distribution corresponding to a quasi-sinusoidal fluctuation of amplitude ΔZ, the critical field has a maximum of order $\lambda \Delta Z / I_0$ when the wavelength is of the order of the

effective boundary thickness. Arguments have been given which suggest that, for ordinary material, whereas the initial susceptibility is mainly dependent on 90° boundary movements, the coercivity corresponds to an appropriate average of the critical fields for 180° boundary movements.

In this necessarily brief indication of the character of the various theories, it is convenient to consider, for illustrative purposes, the quantitative expressions, such as (4.7), obtained for the critical field, H_0, associated with 180° boundary movements. The attraction of the strain theory, however, lay in the fact that it appeared to offer the possibility of interpreting not one, but all the ordinary low field characteristics of ferromagnetics, in a reasonably simple and inherently plausible way, in terms of the internal stress distributions in the material. The comprehensive survey of the theory by Becker and Döring (BD, pp. 147–167, 176–218) gives the impression that, although many matters of detail called for further investigation, the essential basis for a satisfactory physical explanation of the major characteristics had been established. Even if the general scheme were not open to more serious criticism, however, there are two features which detract from its value. The first is that the internal stress parameters (such as ΔZ in (4.7)) which enter into the 'theoretical' expressions for different magnetic characteristics, and which are usually denoted by σ_i, may have a different physical significance in the different expressions, and the relation between them is generally somewhat uncertain. (For a recent discussion of certain aspects of this problem, see Köster (1948).) The coordination of different characteristics is therefore less complete than it at first appears to be, for, in general, the various characteristics neither lead to, nor can they be derived from, a unique and clearly specifiable internal stress distribution. The second point is that little progress has been made in obtaining precise quantitative estimates of the internal stress distributions in ferromagnetic (or other) materials by more direct methods, so that only the roughest independent check can be made of the magnetic estimates. Even rough independent estimates of the probable magnitudes of internal stresses, however, provide some indication of limitations to the range of applicability of the strain theory. For example, insertion of the approximate values for iron of I_0 $(1{\cdot}72\times10^3)$ and $|\lambda|$ (2×10^{-5}) in (4.7), shows that the minimum value of ΔZ (i.e. for $p_0=1$) expressed in kg. mm^{-2} is within 20% of H_0, expressed in oersteds. Now it is hardly reasonable to suppose that variations of internal stress over distances of the order of 10^{-6} cm. (see Table 3.2) from zero to above the normal breaking stress can occur; and it seems clear on this ground alone that some other physical explanation than that given by the strain theory must be sought for coercivities of several hundred oersteds such as are obtained in permanent-magnet alloys. An entirely different mechanism which may here be involved is considered below (§5 (iv)). In many alloys of a more common type, containing, as additions to the ferromagnetic metal, non-ferromagnetic or less ferromagnetic constituents, the coercivity is often markedly increased when the limit of solid solubility is exceeded. It can hardly be denied that this might be due to stresses set up by the inclusions. In many cases, however, the stresses which would be required in annealed materials seem unduly high, and a more direct mechanism suggested by Kersten (1943), which does not involve stress effects at all, seems at first sight much more plausible.

The inclusion theory. The basic idea of Kersten's *Fremdkörpertheorie* is very simple. In a ferromagnetic containing non-ferromagnetic inclusions, a boundary

between domains of the matrix may pass through or include within itself 'impurity' particles which do not contribute to the boundary energy. For a given gross area of boundary the energy will depend on the area occupied by the particles, being a maximum when no particles are included. By idealizing the impurity particles as spherical and uniformly distributed on a cubic lattice, estimates may be made of the increase of boundary energy in the movement of the boundary from a zero field equilibrium position (in which the central plane of the boundary passes through a maximum number of particles), and expressions obtained both for the contribution to the initial susceptibility and for the critical field, in terms of the volume concentration, v_i, of the undissolved non-ferromagnetic impurity, and the ratio of the mean particle diameter, d, to the effective thickness, δ, of the boundary wall. As an approximate expression for the coercivity (obtained from an expression for H_0 by multiplication by a numerical 'averaging' factor), Kersten gives

$$H_c = 2{\cdot}5p(K/I_0)v_i^{2/3}, \qquad \ldots\ldots(4.8)$$

where, for iron and nickel, $K = K_4/3$ (cf. (2.4)), and

$$p = \frac{2(\delta/d)}{1+2(\delta/d)^2}, \qquad \ldots\ldots(4.9)$$

the factor p arising in a similar way to p_0 in (4.7). It is assumed that any local stress anisotropy is negligible in comparison with the crystal anisotropy. The illustrative relation (4.8) is only one of many that are derived, for cubical and rod-shaped as well as spherical particles. In later work (1948) refinements are made in some of the earlier rough calculations (though without any considerable change in the numerical coefficients in the derived relations), and the behaviour of 90° boundaries is examined. In a fuller presentation of the inclusion theory and its applications, published as a book (1943), in which is included an excellent survey of the strain theory, extensive data are collected on coercivities, more particularly of the technically important range of annealed iron alloys containing amounts of other elements which are usually relatively small but above the solubility limit (e.g. C, N, S, Cu), and the data are discussed from the new point of view. The relation (4.8) of the inclusion theory is more nearly amenable to experimental test than (4.7) of the strain theory, for the value of v_i (in contrast to that of ΔZ) is usually known (or determinable) with accuracy. The observed values of H_c can then be compared with the value of $(H_c)_{max}$ which, for spherical inclusions from (4.8) and (4.9), is given by

$$(H_c)_{max} \simeq 1{\cdot}8(K/I_0)v_i^{2/3}, \qquad \ldots\ldots(4.10)$$

and it may fairly be said that the very large number of results examined are at least compatible with the theory. Moreover, although the direct determination of the sizes (and shapes) of inclusions, such as is required for the independent estimate of the factor p in (4.8), would be very difficult, in favourable cases it is much more within the bounds of possibility than a determination of the extent of stress inhomogeneities. In view of such considerations, and of the elegant simplicity and wide field of application of the inclusion theory, it is not surprising that it should have been put forward by its author with so much enthusiasm. Indeed, a much more detailed account of this work of Kersten would be appropriate here, were it not that certain fundamental defects of both the strain and

inclusion theories have been brought to light by critical investigations of Néel (1946 a, b, 1947 a). Although, in the light of these criticisms, a more satisfactory reformulation of the strain and inclusion theories will no doubt be possible, it seems likely that they will then form rather contributions to, or special aspects of, a more general theory, of which a first sketch has been given by Néel.

The disperse field theory. The first criticism made of Kersten's inclusion theory by Néel (1944 d) is that in it the magnetostatic energy associated with inclusions is completely neglected. For a spherical non-ferromagnetic inclusion (effectively a hole) of diameter d and volume v, the difference between the minimum boundary energy (when the central plane of the boundary passes centrally through the inclusion) and the normal energy (when the boundary does not pass through the inclusion) according to Kersten's treatment (for $\delta < d$) is, with sufficient approximation, $\frac{1}{4}\pi d^2\gamma$. The difference in the magnetostatic energy associated with the hole for these two positions of the boundary, calculated exactly by Néel, is, with close approximation, $\frac{1}{2}(\frac{2}{3}\pi I_0^2)v$, or about $0{\cdot}06\pi^2 I_0^2 d^3$. For inclusions in iron the ratio of the magnetostatic to the boundary energy difference is of the order $10^6 d$, which exceeds 100 for some of the larger particle sizes estimated as giving the best agreement with the Kersten relations. Since the Kersten relations, such as (4.8), have been obtained by neglecting an energy effect which may be very much larger than the effect taken into account, it is clear that the whole theoretical treatment of inclusions requires reconsideration. In this same paper Néel shows that the magnetostatic energy may be reduced by the formation of appropriate secondary domain structures, as has already been mentioned (§ 3 (ii)), and the general character of the powder patterns round microscopically visible inclusions (Figure 3.12) is in agreement with his predictions. Taking this effect into account, Néel outlines a modified treatment of the coercivity arising from inclusions, but this particular treatment, as he subsequently states (1947 a), is shown to be erroneous by his later work.

In his most comprehensive paper on this subject, dealing with the principles of a new general theory of the coercive field, Néel (1947 a) first puts forward a criticism of the quantitative treatment adopted in developing both the strain and inclusion theories. This criticism, as quantitatively formulated, makes it probable that, except perhaps in some very special cases, the formulae derived are invalid, or at least inapplicable, and that the mechanisms proposed could account for coercivity values only a small fraction of those observed. The method usually, if not invariably, adopted in treating the effect of stress variations on the variation of boundary energy with position is to assume, for simplicity, that the stress varies in the direction of movement of the boundary (e.g. in the x direction in Figure 4.1) but that it is constant over the plane of the boundary (e.g. in the yz plane). Any subsequent averaging process does not affect the initial assumption of a one- rather than a three-dimensional stress variation over volumes comparable in size with those traversed by a moving boundary. Similarly, in the inclusion theory, the inclusions are assumed to be uniformly distributed on the points of a lattice, the planes of which are parallel to the boundary planes considered. Once it has been pointed out, it is obvious that a purely random distribution of local stresses or inclusions can give rise only to a very much smaller positional variation of boundary energy. From a quantitative treatment of the problem Néel concludes that the maximum possible variations of boundary energy could account for coercivities of the order of only a tenth of an oersted if the boundary

is regarded as strictly plane, and of a few oersteds if allowance is made for the possibility of deformation of the boundary. It is quite impossible to account for observed coercivities of tens and even hundreds of oersteds.

Although internal stresses and inclusions may have a negligible direct effect on boundary energy variation, there is no doubt that they are responsible, in some way, for the diversity in the low field properties of different materials, or, more specifically, that they give rise indirectly to a variation of energy with boundary position. Néel suggests that the randomly distributed localized irregularities result in fluctuations in the direction and the magnitude of the intrinsic magnetization, as indeed they must do. The magnetization is therefore not strictly uniform, even within a single domain, and the fluctuations give rise to non-zero values of the divergence of I; free poles are developed and corresponding magnetic field distributions. The energy associated with these disperse fields (*champs magnétiques de dispersion*) varies with the position of domain boundaries. The variable energy may be regarded as arising from the mutual potential energy of the free pole regions, islands of positive or negative magnetic charge. The sign of the charge depends on the general direction of magnetization; thus an island which is positive when the magnetization is in one direction (e.g. I_0 upwards in Figure 4.1) would become negative when, through boundary movement, the general magnetization direction is reversed (e.g. I_0 downwards); and an island 'bisected' by a boundary will be oppositely charged on the two sides. Thus the volume energy, arising from the mutual potential energy of the magnetically charged islands in any region varies with the position of boundaries in that region. Indirectly, internal stresses and inclusions can therefore give rise to a variation of energy with boundary position similar to that shown in Figure 4.1. Moreover since the magnetostatic energy associated with fluctuations in the intensity of magnetization may greatly exceed the directly produced boundary energy change (as has been noticed in connection with the first criticism of the inclusion theory), it is reasonable to expect that the disperse field effect, even with a completely random distribution of stresses or inclusions, may give rise to energy variations of the order of magnitude required to account for observed values of coercivity and other low field characteristics.

The quantitative mathematical development of the ideas which have been outlined obviously presents great difficulty, and Néel's derivation of even approximate expressions for the coercivity is a remarkable feat. Denoting by C the deformation energy affecting a fractional volume v_s of the material, for which the magneto-crystalline anisotropy coefficient is K, Néel obtains, for $C/K \leqslant 1$,

$$H_c = \frac{4}{15\pi} \frac{v_s C^2}{K I_0} \left\{ 1{\cdot}386 + \tfrac{1}{2} \ln \left(\frac{2\pi I_0^2}{K} \right) \right\}, \qquad \ldots\ldots(4.11)$$

and for $K/C \ll 1$, $$H_c = 0{\cdot}69 \frac{v_s C}{I_0} \left\{ 1{\cdot}386 + \tfrac{1}{2} \ln \left(\frac{6{\cdot}8 I_0^2}{C} \right) \right\}. \qquad \ldots\ldots(4.12)$$

For non-magnetic inclusions of fractional volume v_i, the relation derived is

$$H_c = \frac{2K v_i}{\pi I_m} \left\{ 0{\cdot}386 + \tfrac{1}{2} \ln \left(\frac{2\pi I_m^2}{K} \right) \right\}, \qquad \ldots\ldots(4.13)$$

where I_m is the mean value of the spontaneous magnetization. The relation between these expressions and (4.7) and (4.8) may be noted. The numerical significance of the relations becomes clearer on taking particular examples

given by Néel. Assuming that the internal stresses are tensions, Z_i, orientated at random, and of magnitude 30 kg. mm^{-2}, affecting a fractional volume v_s, then, putting $C=3\lambda Z_i/2$, and using the appropriate values for I_0 and K, the formula obtained for iron is

$$H_c=2{\cdot}1v_s+360v_i, \qquad \ldots\ldots(4.14)$$

and for nickel,

$$H_c=330v_s+97v_i. \qquad \ldots\ldots(4.15)$$

These two relations indicate that in iron the effect of inclusions is predominant, in nickel that of internal stresses. For comparison with experiment, Néel uses the data collected by Kersten for iron alloys. On a logarithmic plot of H_c against v_i the points are very scattered, but the line given by (4.14) drives centrally through them in a most convincing manner.

It is not possible to give here a satisfactory account of the detailed argument by which the relations (4.11) to (4.13) are derived. For this, reference may be made to the original paper, and to a later short outline given by Néel (1949). The mathematical treatment of so complex a physical problem is itself unavoidably complicated (e.g. in the manipulation of the Fourier representations of randomly distributed stress variations), but, apart from this, it is not unfair to say that obscurities in the presentation do not seem to arise wholly from the inherent complexities of the problem. It is important to bear in mind that the validity of a mathematical (or physico-mathematical) argument in itself cannot be confirmed merely by an agreement with experiment of approximate relations which may have been derived; and that the value of a theory as interpretative of observable phenomena cannot be properly assessed until the essential details in the argument from the premises to the conclusions have been clearly presented. The position may be expressed, loosely and incautiously, as follows. The 'conclusions' of the strain, inclusion and disperse field theories all agree (more or less) with experiment. With the first two, the deductive arguments are clear and satisfying; but the premises are wrong, or at least inadequate, as a representation of the physical situation. With the disperse field theory, the premises may be tentatively accepted as correct, or at least more nearly adequate; but before the theory can be regarded as 'established', a clearer development of the deductive arguments is required, together with a wider and more searching comparison of the theoretical conclusions with experiment.

Little useful purpose would be served by dealing here with other aspects of boundary movement in relation to magnetic characteristics, as most of these (e.g. the reversible movement in relation particularly to initial susceptibility) have so far been treated only from the standpoint of the older theories. Some of the relations obtained (e.g. that between maximum initial susceptibility and magnetostriction, cf. I, p. 65, equation (2.22)) may still be found to be applicable, though perhaps over a more limited range, while others will undoubtedly require considerable modification in interpretation, if not in quantitative expression. The particular need is for a clearer and more extended development of the disperse field theory. The more general need is for a linking together of what may be called the large scale characteristics of domain boundaries, such as the formation of regular equilibrium structures dependent on field, as indicated by theory (§ 2 (iii), § 3 (i)) and manifested in the powder pattern experiments (§ 3 (ii)), with the small scale characteristics which underlie the Barkhausen effect (§ 3 (iii)), and with which this subsection (4 (ii)), in so far as irreversible movements have been

considered, has been mainly concerned. It may well happen that substantial advances will have been made in both these directions by the time this Report is in print.

4(iii). *Reversible Rotational Processes*

Rotational processes in single crystals have already been fully discussed (§ 2). The main question to be considered here is the relation between the variation with field of the magnetization of a polycrystalline aggregate, over a rotational range, and that of single crystals. When the field is increased from zero, with the specimen initially in a demagnetized state, the changes in magnetization in the lower fields are mainly due to boundary movements. In higher fields rotational changes predominate, and the range to which most attention has been devoted in attempting to correlate the behaviour of single crystals and polycrystalline materials is that of approach to saturation. In a field decreasing from a high value, the predominance of rotational changes persists over a more extended range. In general, there is no good reason to suppose that the change in magnetization over the range from saturation to remanence is wholly rotational (or even wholly reversible); it is none the less of interest to compare observed values of the remanence with values which can, in some cases, be derived theoretically on the assumption that only rotational processes are involved. There is little doubt that in most materials the rotational processes which take place are entirely reversible, and only such reversible processes are here under consideration; the probability of the occurrence of irreversible rotational processes in certain types of high coercivity materials is discussed below (§ 4(iv)).

Approach to saturation. Empirically, in the range in which I/I_0 approaches unity, the variation of magnetization with field may be represented by

$$I = I_0\left(1 - \frac{a}{H} - \frac{b}{H^2}\right) + cH \qquad \ldots\ldots(4.16)$$

(cf. I, p. 61, BD, p. 167.) The term in H, which may be attributed to the increase in the intrinsic magnetization above the spontaneous magnetization value at the temperature, is very small at temperatures well below the Curie point, and need not be considered here, though it is important in connection with magneto-thermal effects (§ 4(v)). The term in H^{-1} is usually dominant in fields above a few thousand oersteds, as shown particularly by the work of Weiss and his collaborators, and a is referred to as the coefficient of magnetic hardness (*dureté*). None of the earlier suggestions as to the origin of this term is satisfactory (cf. BD, pp. 175–6); an explanation recently proposed by Néel (1948 a) is outlined below. The term in H^{-2} may be connected with rotation of the magnetization vectors towards the applied field direction against the local anisotropy fields. This process gives rise to a variation of magnetization with field which may be quite generally represented by

$$1 - (I/I_0) = \sum_{n=2} b_n H^{-n}. \qquad \ldots\ldots(4.17)$$

The leading coefficients in this series have been calculated by Akulov (1931), and more completely by Gans (1932), for an aggregate of quasi-independent single crystals orientated at random. For iron and nickel the theoretical value of the first coefficient so obtained is given by

$$b' = b_2 = \frac{8}{105}(K_4/I_0)^2. \qquad \ldots\ldots(4.18)$$

Measurements on both iron and nickel in the form of polycrystalline wires by Czerlinski (1932), and more recently on nickel, over a wide temperature range, by Polley (1939), yielded values for b which, putting $b=b'$, gave values of K_4 in excellent agreement with those obtained from measurements on single crystals (cf. BD, pp. 171–5).

This agreement is very remarkable for, as seems first to have been pointed out by Holstein and Primakoff (1941), in the derivation of (4.18) the mutual magnetic interaction of the crystalline grains is neglected. The field acting on any particular grain owing to the difference in the directions of magnetization of surrounding grains is likely to be comparable with the applied field, when this is in the usual range of a few hundred to a few thousand oersteds; the value of b given by (4.18) should be approached only for $(H/4\pi I_0)\gg 1$. The detailed calculation of b (or more generally of I/I_0) as a function of $H/4\pi I_0$, the relevant parameter, is complicated, and it is very satisfactory that Néel (1948 b), in a later but completely independent treatment of the problem, obtains the same result as Holstein and Primakoff. Retaining b' for the theoretical value of the coefficient of H^{-2} given by (4.18), using b'' for the new value, and writing

$$\alpha = H/4\pi I_0, \qquad \ldots\ldots(4.19)$$

the result may be expressed by

$$b''/b' = f(\alpha) = \tfrac{1}{2}\left[1+\frac{\alpha}{2(1+\alpha)}+\frac{\alpha^2}{4(1+\alpha)^{3/2}}\ln\left\{\frac{(1+\alpha)^{1/2}+1}{(1+\alpha)^{1/2}-1}\right\}\right]. \qquad \ldots\ldots(4.20)$$

For $\alpha \ll 1$ (i.e. $H \ll 4\pi I_0$), $\quad f(\alpha)\to\tfrac{1}{2}$;

for $\alpha \gg 1$ (i.e. $H \gg 4\pi I_0$), $\quad f(\alpha)\to 1$.

It may be noticed that for small values of α, $f(\alpha)$ varies slowly. For $\alpha=1$ (corresponding to values of H of about 6,000 and 22,000 for nickel and iron), $f(\alpha)=0{\cdot}7$, still well below the value unity corresponding to the theoretical value when magnetic interactions are neglected. To test the new relations measurements were made by Néel on a thin wire of pure nickel subjected to a thorough annealing treatment in carefully purified hydrogen. An almost exactly linear variation of I with $1/H^2$ was found over a field range from about 150 to 750 oersteds. The value found for b was little more than half that given by Polley's results (290 as compared with 550); over the range covered the mean value of $f(\alpha)$ is about 0·515, so that, with the new formulae, there is entirely satisfactory agreement between the value of K_4 derived from measurements on polycrystalline material and that obtained directly. The higher values of b found earlier (which, with an erroneous theoretical formula, fortuitously gave satisfactory values of K_4) are tentatively attributed by Néel to residual strains or other imperfections in the specimens used.

In a related paper dealing with the a/H term, and also describing work virtually completed in 1945, Néel (1948 a) points out that no mechanism involving perturbing forces can give rise to a rigorous law of this type, which implies an indefinitely large work of magnetization. This rules out most of the suggestions that have been made, including those of Brown (1940 b, 1941), whose study of the effect of point, line and plane concentrations of force, or, less formally, of dislocations, is none the less of great interest. The observed temperature variation of a,

moreover, seems to be incompatible with an at first sight promising type of explanation in terms of the statistical equilibrium of small domains, treated as molecules of large permanent moment (Stoner 1936). Néel examines both theoretically, and, with Lorin, experimentally, another possibility: that the effect may arise from the presence of cavities or non-ferromagnetic inclusions. The average state of magnetization is determined by the applied field and the internal fields due to the cavities. As in his development of a general theory of coercivity (§4(ii)), and the treatment of the effect of the magnetic interaction between crystalline grains on the approach to saturation, Néel makes ingenious use of the Fourier representation of an irregular distribution of spontaneous magnetization. Writing v for the volume concentration of cavities (or non-ferromagnetic inclusions) I_m for saturation magnetization of the 'porous' aggregate (so that $I_m=(1-v)I_0$, where I_0 is the saturation magnetization of the ferromagnetic itself), and putting

$$\alpha = H/4\pi I_m, \qquad \ldots\ldots (4.21)$$

Néel's final expression for the approach to saturation may be put in the form

$$1-(I/I_m)=\frac{v}{2(1-v)}\left[\frac{2+3\alpha}{4(1+\alpha)^{1/2}}\ln\left\{\frac{(1+\alpha)^{1/2}+1}{(1+\alpha)^{1/2}-1}\right\}-\frac{3}{2}\right]. \qquad \ldots\ldots (4.22)$$

For large values of α, the limiting form of the expression in square brackets is $(2/15)\alpha^{-2}$, so that eventually the approach to saturation is linear in $1/H^2$. For iron, however, this would hold only in very high fields, and over a range from a few thousand to fifty thousand oersteds the variation is nearly linear in $1/H$. Measurements were made on specimens of porous iron, prepared from iron powder, and of densities ranging from about 7·2 to 4·9. For fields above about 1,000 oersteds the agreement with the theoretical expression is extremely good. It may be noted that over the wide range nearly linear in $1/H$ the coefficient a increased with decreasing density from about 50 to 300. There can be little doubt that the long-standing problem presented by the a/H term first given by Weiss as long ago as 1910 in an empirical expression for the approach to saturation has at last, in principle, been satisfactorily resolved. This does not mean that a physical explanation of the magnitude of the hardness coefficient, a, found for any particular material can immediately be given; but the direction in which to look is now clear.

Remanence. On the implicit assumptions that magnetic interaction between the individual crystals may be neglected, and that the change from saturation to remanence occurs only by reversible rotational processes, Gans calculated the remanence of random aggregates of crystals of the iron, nickel and cobalt types in the paper (1932) to which reference has been made above. Writing j_r for I_r/I_0, the results, which do not appear to be widely known, are as follows:

(*a*) For cubic crystals with cube edges as easy directions (e.g. iron),

$$j_r=(3/\pi)2^{1/2}\tan^{-1}2^{-1/2}=0{\cdot}8312.$$

(*b*) For cubic crystals with cube diagonals as easy directions (e.g. nickel),

$$j_r=3^{1/2}/2=0{\cdot}8660.$$

(*c*) For hexagonal crystals with the hexagonal axis as the easy direction (e.g. cobalt),

$$j_r=0{\cdot}5.$$

The value under (*c*), $j_r = \frac{1}{2}$, would apply quite generally for randomly orientated crystals with a single easy axis of magnetization, such as might arise even in cubic crystals as a result of stresses. For aggregates with preferential orientations the reduced remanence might have any value from zero to unity, as illustrated by the curves of Figure 4.3. For unstrained cubic crystals, with K_4 positive, the values of the remanence for special orientations are indicated by Figure 2.1 (see also I, pp. 57, 59).

For materials under applied stress of such magnitude that λZ exceeds K the experimental results are, on the whole, in good accordance with the simple theory. For example, the value of j_r for a wire under increasing tension approaches unity if λ is positive, zero if λ is negative. For annealed materials in which the state of random orientation of the individual crystals would be expected to be closely approached the j_r values differ widely. For both iron and nickel, from data collected by Bozorth (1948), the j_r values range from about 0·3 to 0·7. The values are almost invariably lower than those given under (*a*) and (*b*) above, and are usually lower than 0·5.

The effect of magnetic interactions on remanence does not appear to have been explicitly discussed, though the curve and figures given by Néel (1947 a, pp. 325, 329) indicate, in place of the j_r value of 0·5, as given under (*c*), the higher value 0·7. Bozorth, on the other hand, suggests that the demagnetizing field of any single domain will tend to decrease the magnetization in it by a process involving reversal of part of the domain and the formation of a boundary wall. Such a process would clearly be favoured by small values of the crystal or strain anisotropy, and Bozorth sees here a possible explanation of the very low remanence of iron–nickel and other alloys which have been well annealed and have low crystal anisotropy. It is, of course, far from clear under what circumstances the 'negative' self-demagnetizing field of a domain would outweigh the intercrystalline field, which the treatment of Holstein and Primakoff and of Néel indicates to be positive and large.

From an experimental standpoint the inadequacy of a treatment based on the assumption that only reversible rotational processes occur from saturation to remanence is not surprising. In general the descending curve is not closely retraced when the field is increased from zero (in other words, the irreversible susceptibility is an appreciable fraction of the differential susceptibility) and, connected with this, the Barkhausen effect becomes marked well before remanence is reached on the descending curve. This is clearly illustrated, for a fairly representative material, in the curves of Figure 3.18. Reversible boundary movement may, of course, also occur, and it should be mentioned that a tentative theoretical treatment has been given of the relation between the initial susceptibility and the reversible susceptibility at remanence assuming that only 90° boundaries are involved, though no clear-cut conclusions can be drawn from such comparisons with experiment as have been made (BD, pp. 160–62).

The analysis of a magnetization curve in the remanence region presents a very complex problem. A theoretical treatment based on the assumption of pure rotations does not provide a complete answer to the question of just what is taking place inside the material; but it does provide a sufficient part of the answer to point the way clearly to some of the further experimental and theoretical investigations which are required.

4 (iv). *Rotational Processes and Hysteresis*

In materials which have been developed for permanent magnets (§ 5 (iv)), coercivities exceeding 500 oersteds have been obtained with values of the remanence B_r exceeding 10^4 gauss. Such a combination of characteristics could not be accounted for in terms of domain boundary movements on the basis of the Becker–Kersten theory (nor, it may be added, on the basis of Néel's much more general disperse field theory, though this can account satisfactorily for much higher coercivities than can the earlier strain theory), and a different type of mechanism seemed to be indicated. Investigation of the alloy of approximate composition $Fe_2Ni\,Al$, which is the prototype of an important group of permanent magnet alloys, by x-ray and other methods (I, pp. 107–8) showed that high coercivity is developed by a heat treatment intermediate between that required to retain the high temperature single-phase state (quenching) and that required to give a complete breakdown into a two-phase state (very slow cooling). The suggestion made by Bradley and Taylor (1938 a, b), and generally accepted, was that the alloy then contained iron-rich 'islands', too small to be regarded as a precipitated phase in the usual sense, and the high coercivity was tentatively attributed to immense local stresses set up. In view of the suspected inadequacy of the stress interpretation, it appeared desirable to carry out a quantitative theoretical investigation of the magnetic behaviour of small particles of a ferromagnetic embedded in a non-ferromagnetic matrix. In such particles, if they are below a certain critical size, depending on shape, no domain boundaries could be formed (§ 3 (iv)); they would be single-domain particles, uniformly magnetized, and change of direction of the magnetization vector could occur only by the rotational process. Preliminary calculations, made in 1939, showed that very high coercivities could arise in this way in an aggregate of single-domain, magnetically anisotropic particles. The magnitudes of these coercivities are comparable with the values earlier estimated as upper limits with magneto-crystalline or strain anisotropy, but which were regarded as physically unrealizable owing to boundary movement effects (BD, p. 110). With shape anisotropy even higher values are possible. It was not until much later that a detailed systematic investigation of the general problem could be made (Stoner and Wohlfarth 1947, 1948). In the interim, a considerable amount of work had been carried out in France on the development of iron powder materials for permanent magnets, in which Néel and his collaborators played a large part. The theoretical considerations which served as a guide in this work, as outlined in three short papers published later by Néel (1947 b, c, d), were essentially the same as those indicated above, and, although no elaborate numerical calculations appear to have been undertaken, the estimated values for certain key quantities are in good agreement with those obtained in the more detailed treatment. Further reference is made to the powder magnet work below (§ 5 (iv)). Investigations by Guillaud had also shown that high coercivities could be obtained with finely divided ferromagnetic materials, and that, in the materials he studied, the effects could be attributed to single-domain particles with anisotropy primarily of magneto-crystalline origin. Guillaud's work in this field was not widely known, and even now no full account of it is generally available; but in a recent note Guillaud (1949) outlines what had been done, and gives references to his earlier publications. The basic ideas arise naturally from the outlook which was developing in 1939, and may well

have occurred to others than those mentioned. The following outline of the theoretical treatment of rotational processes in single-domain particles and of the physical bearing of the results is based on the paper of Stoner and Wohlfarth (1948), which should be consulted for further details.

Assuming first that the magnetic anisotropy is due purely to shape, it is convenient to consider a single-domain uniformly magnetized particle in the form of a prolate spheroid, for which it may be shown that I_0 lies in the plane defined by H and the polar axis. The relevant part of the total energy E, equivalent here to the free energy F, is then the sum of the demagnetizing field energy E_D and the external field energy E_H. With the symbols of Figure 4.2,

$$E = E_D + E_H = \tfrac{1}{2} I_0^2 (N_a \cos^2\psi + N_b \sin^2\psi) - H I_0 \cos\phi, \quad \ldots\ldots (4.23)$$

where N_a and N_b are the demagnetization coefficients along the polar and equatorial axes.

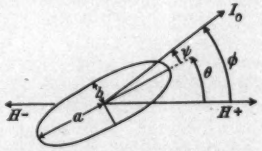


Figure 4.2. Symbols for prolate spheroid in magnetic field.

After some manipulation, an expression for the variable part of a suitably reduced energy may be obtained in the form

$$\eta = -\tfrac{1}{4}\cos 2\psi - h\cos\phi, \quad \ldots\ldots (4.24)$$

where

$$\eta = \frac{E}{(N_b - N_a) I_0^2} - \frac{1}{4}\frac{N_b + N_a}{N_b - N_a},$$

and

$$h = H/\{(N_b - N_a) I_0\}. \quad \ldots\ldots (4.25)$$

Information is first required as to how the angle, ϕ, between the magnetization vector, I_0, and the positive direction of the field (or, more usefully, the reduced resolved magnetization, I_H/I_0, equal to $\cos\phi$) varies with the field H (or the reduced field h) for a given orientation, θ, of the spheroid in the field. It is better, therefore, to rewrite (4.24) as

$$\eta = -\tfrac{1}{4}\cos 2(\phi - \theta) - h\cos\phi. \quad \ldots\ldots (4.26)$$

For given values of θ and h, stationary values of η are given by

$$\partial\eta/\partial\phi = \tfrac{1}{2}\sin 2(\phi - \theta) + h\sin\phi = 0, \quad \ldots\ldots (4.27)$$

and these correspond to minima, points of inflection, or maxima for

$$\partial^2\eta/\partial\phi^2 = \cos 2(\phi - \theta) + h\cos\phi \gtrless 0. \quad \ldots\ldots (4.28)$$

Although the general character of the change of ϕ with h may in part be obvious physically (e.g. as h changes from a large positive value to zero, ϕ changes from 0 to θ) the precise course of the change, and whether or not discontinuities occur, can be determined only by obtaining the solutions of (4.27) which, in the light of (4.28), are physically relevant. Analytical expressions relating h, ϕ and θ can be derived corresponding to critical values of h at which discontinuities in magnetization occur, but in general solutions of the equation (4.27) must be obtained

numerically. Detailed numerical results for an effectively complete range of fields and orientations are given in the original paper. The general character of these results is illustrated in a physically significant form in the curves of Figure 4.3, which, with the caption, is self-explanatory.

The magnetization curve for an assembly of prolate spheroids orientated at random is shown in Figure 4.4. The integrations involved in obtaining precise numerical values for the mean magnetization for values of h from $-0{\cdot}5$ to $-1{\cdot}0$ on

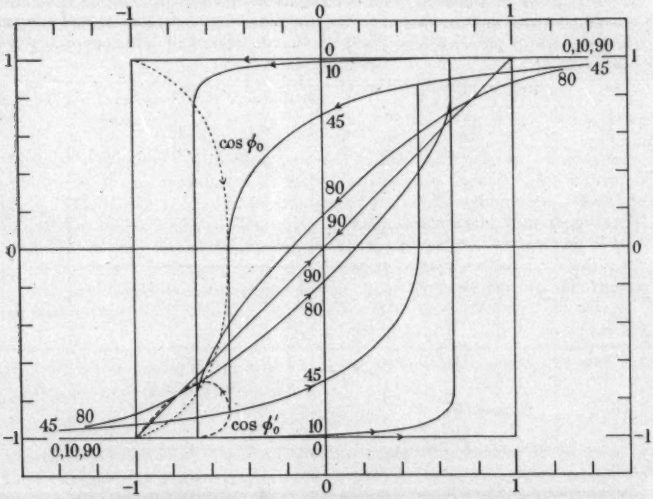


Figure 4.3. Magnetization curves for uniformly magnetized prolate spheroids.

The resolved magnetization in the positive field direction is $I_0 \cos\phi$. The field H is given by $H=(N_b-N_a)I_0h$. The angle, θ, between the polar axis and the field is given, in degrees, on the curves. The dotted curves give $\cos\phi_0$ and $\cos\phi_0'$, where ϕ_0 and ϕ_0' are the angles made with the positive field direction by the magnetization vector at the beginning and end of the discontinuous change at the critical value, h_0, of the field. (Stoner and Wohlfarth 1948.)

(Reproduced, by kind permission, from the *Philosophical Transactions of the Royal Society*.)

the descending curve are very troublesome. In drawing the curve of Figure 4.4 the original numerical values, at intervals of $0{\cdot}1$ in h, have been supplemented by further values, due to Rhodes (1948), at intervals of $0{\cdot}02$ over the range from $-0{\cdot}5$ to $-0{\cdot}6$; the discontinuity in the slope of the curve at $h=-0{\cdot}5$ is clearly shown. It is to be noted that there are no discontinuous changes on the descending (or ascending) branch of the curve until $|h|$ exceeds the 'mean coercivity' value, $|h_c|\simeq 0{\cdot}479$. In Figure 4.4 the magnetization curve for an assembly of oblate spheroids is also shown. For ellipsoids in the special form of oblate spheroids there is no hysteresis, but a discontinuity in magnetization occurs at $h=0$. As the field decreases through zero, the magnetization vector can leave the plane defined by the field and the polar axis, and rotate through 180° in the

equatorial plane from the equilibrium direction appropriate to a small positive field to that appropriate to a small negative field. For the general ellipsoid the relations, which have been considered in less numerical detail, are very complicated, but it is clear that the mean magnetization curves would be intermediate between those for prolate and oblate spheroids.

The initial curve for the prolate spheroid assembly shown in Figure 4.4 is the mean of the ascending and descending branches of the hysteresis curve. It corresponds to a definite initial demagnetized state in which, for spheroids orientated within any angular range, the magnetization vector for one half is along the polar axis in one sense (e.g. $\phi=\theta$), and for the other half in the opposite sense ($\phi=\pi+\theta$).

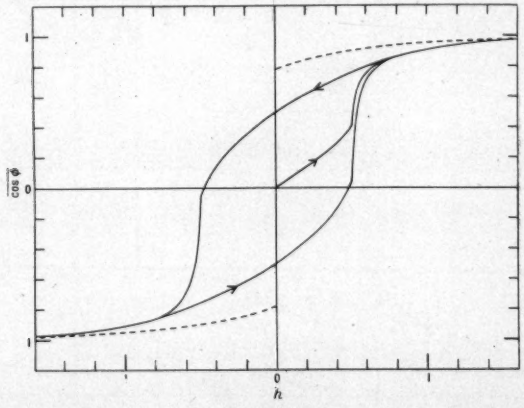


Figure 4.4. Magnetization curves for prolate (full curves) and oblate (broken curves) spheroids orientated at random.

The mean resolved magnetization per spheroid in the positive field direction, or the resultant magnetization in this direction of the assembly, is proportional to $\overline{\cos\phi}$. $H=(\lfloor N_a-N_b \rfloor)I_0 h$.

(*After* Stoner and Wohlfarth 1948, and Rhodes 1948.)

If the anisotropy of the particles is due entirely to shape, then, from (4.25), the maximum coercivity, corresponding to $h=1$ (see Figure 4.3), for prolate spheroidal particles orientated parallel to the field is equal to $(N_b-N_a)I_0$. For random orientation of the particles, the 'mean coercivity', $h_c=0{\cdot}479$, corresponds to a value of H_c a little under half as great. The upper limit of N_b-N_a for needle-like particles is 2π. Thus the upper limit to coercivity due to shape anisotropy is

$$(H_c)_{\max}=2\pi I_c, \qquad \ldots\ldots \quad (4.29)$$

giving values of about 11,000, 3,000 and 9,000 for iron, nickel and cobalt respectively. These are extreme values, but the values are high even for comparatively slight deviations from a strict spherical form; for a dimensional ratio as low as 1·2, for example, the corresponding values of $(H_c)_{\max}$ are about 1,530, 430 and 1,260. It should be noted that if the particles are embedded in a non-ferromagnetic matrix, I_0 is the saturation intensity for the material of the particles,

not the mean value for the aggregate. The field, in all the relations, is that 'applied to' the particle, and this may differ significantly from the field acting on the specimen containing the particles, whether before or after 'correction' for the demagnetizing field of the specimen as a whole; in applications it is essential that this should be taken into account (cf. Néel 1947 c, Weil 1947).

The basic equation (4.27) has been derived with specific reference to shape anisotropy. Precisely the same equation is obtained for spherical particles subject to tension (for λ positive) or compression (λ negative), and for which the magneto-strain anisotropy is of the type corresponding to the energy expression (cf. (3.19), (4.1))

$$E_Z = \tfrac{3}{2}\lambda Z \sin^2\psi, \qquad \ldots\ldots(4.30)$$

the relation between H and h then being

$$H = (3\lambda Z/I_0)h. \qquad \ldots\ldots(4.31)$$

Without considering the physical question of how such magneto-strain effects might arise, it may be noted that, with Z expressed in kg.mm^{-2}, the values of $(H_c)_{max}/Z$ are about 3, 30 and 4 for iron, nickel and cobalt respectively. With the mechanism under consideration very high coercivities could, therefore, hardly be accounted for primarily in terms of magneto-strain effects for particles of iron or cobalt; these effects may, however, be of great importance for nickel particles.

Magneto-crystalline anisotropy of the type corresponding to the energy expression (cf. (2.2))

$$E_K = K \sin^2\psi \qquad \ldots\ldots(4.32)$$

again leads to equation (4.27), with the same consequential results, but with the relation between H and h now given by

$$H = (2K/I_0)h. \qquad \ldots\ldots(4.33)$$

For cubic crystals with magneto-crystalline energy expressions of the type of (2.1) the field-orientation–resolved magnetization relations, which would be much more complex than when the anisotropy is of the uniaxial type, have not been investigated in detail. For iron, however, for which the anisotropy coefficient K_4 is positive, the maximum coercivity is the same as that derivable from (4.33), namely $(H_c)_{max} = 2K/I_0$. For iron and cobalt the values indicated for $(H_c)_{max}$ are about 500 and 6,000, and for nickel the value is hardly likely to exceed 200.

For aggregates containing iron particles, coercivities appreciably exceeding 500, such as have been obtained with iron powder magnets (Néel 1947 b), must almost certainly be attributed to shape anisotropy. Analyses of Guillaud's results on the compound MnBi by Kittel (1948), and of those on Mn_2Sb by Guillaud (1949), show that here the high coercivities are connected with magneto-crystalline anisotropy. In most cases, however, definite conclusions cannot be drawn as to the type of anisotropy primarily involved purely from a consideration of the magnitudes of the observed coercivities. Further investigations are necessary, among which it is obvious that an examination of the temperature variation of the coercivity, such as has been made by Weil *et al.* (1947, 1948) for some of the powder magnet materials, may be of great value (see § 5 (iv)).

The three types of materials to which the treatment outlined seems to be particularly relevant are, as indicated in the original paper, (*a*) metals and alloys containing ferromagnetic impurities, (*b*) powder magnets, and (*c*) high coercivity

alloys of the dispersion hardening type. It is clearly impossible to discuss the various materials *in extenso* here, though reference will be made to some of them below (§5(ii), (iv)). It may perhaps be appropriate to draw attention to an interesting application of the general ideas to the interpretation of the characteristics of magnetic recording wire (Hobson and Osmond 1948). The most frequently used wire material is chromium–nickel stainless steel which, after the cold working and heat treatment found appropriate, has a coercivity of the order of 500, and contains ferromagnetic ferrite particles dispersed in a non-ferromagnetic austenitic matrix. From a determination of the relative amount of ferrite present, and an examination of the range of undistorted amplitude and frequency response, Hobson and Osmond have made estimates, admittedly rough, of the mean lengths and cross sections of the ferrite particles. The ratio of length to diameter is found to be about 10, and the diameter about 10^{-5} cm. or less. These values are completely consistent with the theoretical estimates (§3(iv)) of the particle sizes below which boundary formation cannot occur. Probably the most extensive field of application of the theoretical treatment is to dispersion hardening alloys, and the survey made of the earlier work on these alloys left little doubt as to the value of the new approach to the interpretation of the magnetic characteristics. Most of these alloys, however, are extremely complex, and, in general, there is little likelihood of being able to explain their magnetic properties in terms of one type of elementary process. The possibilities and limitations of explanations in terms of rotational processes in single-domain particles will become clearly apparent only when much further experimental work has been carried out on these remarkable alloy systems.

It may finally be remarked that the single-domain particle mechanism is not so sharply distinct in character from other mechanisms which have been considered (§4(ii)) as may at first appear. If the particles were larger than those considered, domain boundary formation would occur within them, and the behaviour of the particles, or segregates, would become similar to that of ferromagnetics in bulk. A clear-cut simplicity is given to the theoretical treatment by limiting it to single-domain particles, a limitation which appears to correspond closely to the physical conditions in many types of materials characterized by high coercivity. Sharply defined particles, however, correspond to an extreme form of fluctuation in composition, with the associated fluctuation in magnitude of the intrinsic magnetization, and the disperse fields which are consequentially developed are here the demagnetizing fields of the particles themselves. It may be noticed that the importance of purely magnetostatic effects, the recognition of which has already been mentioned as characterizing so much of the more recent theoretical work on ferromagnetism, is again very evident in the whole of the treatment of hysteresis associated with rotational processes in single-domain particles.

4(v). *Magneto-thermal Effects*

The magneto-thermal effect to be discussed here is the change of temperature accompanying a change of field applied to a ferromagnetic under adiabatic conditions. It is often convenient to consider, instead of the change of temperature, ΔT, the corresponding 'heat developed', $\Delta Q'$, per unit volume of the material, related to ΔT by

$$\Delta Q' = C\rho\Delta T, \qquad \ldots\ldots(4.34)$$

where C is the specific heat and ρ the density. By far the best known of all magneto-thermal effects is the heat developed, commonly known as the hysteresis loss, when a ferromagnetic is taken through a magnetization cycle. This loss may be accurately determined by well-known methods, although the actual temperature change for a single cycle is usually very small, of the order of 10^{-4} deg. c. for soft magnetic materials. It is hardly necessary to discuss hysteresis loss as such here, for although it is, technologically, an important characterizing magnitude for certain types of material, it is clearly not a primary characteristic. It is sufficient to recall that it may be determined from the area of an I, H or B, H loop, and that it corresponds to the aggregate of the irreversible thermal changes as the cycle is traversed. In contrast to the irreversible changes contributing to the hysteresis loss, the changes of temperature with field in high fields are purely reversible. The temperature change may be small at room temperature, but in the Curie point region changes of more than 1° are observed in nickel, and more than 2° in iron for field changes of the order of 10^4 oersted. These temperature changes may be associated with changes in intrinsic magnetization (I, pp. 50, 77, 81), and the effect is essentially similar in origin to that utilized in obtaining very low temperatures by the adiabatic demagnetization of paramagnetic salts. The main experimental investigations are those of Weiss and Forrer (1926) on nickel, and of Potter (1934) on iron, and theoretical discussions of the results are given by these authors and also by Stoner (1936). The high-field effect is not distinct in character from that in lower fields, and it is misleading to refer to it, as is sometimes done, as *the* magneto-caloric effect.

In low and moderate fields the change of temperature with change of field is, like the change of magnetization, far from linear. The difficulty of investigating the adiabatic temperature changes in these fields is that if any detailed information is to be obtained as to the course of the changes as the magnetization curve is traversed, the field changes must be made in small steps; the changes in temperature to be measured are then small fractions of the extreme range of the temperature change, which itself is often much less than 10^{-3} deg. Since the pioneer work of Adelsberger (1927), the experimental methods have been refined, and measurements made on representative materials by Constant (1928), Honda, Okubo and Hirone (1929), Ellwood (1930), Townsend (1935), Okamura (1936) and Hardy and Quimby (1938). More recently comprehensive investigations, surveyed in a lecture by Bates (1949), have been made on iron, nickel and cobalt and a number of alloys by Bates with Weston (1941), Healey (1943), Edmondson (1947) and Harrison (1948 a, b). The experimental arrangements which have been adopted differ in such things as the form of the test sample (single specimen, or built-up multiple specimen), the number and manner of attachment of the thermocouples, the way of connecting them to the galvanometer, and the method of calibration. For convenience and range of applicability a single specimen method is much to be preferred. In the method developed by Bates (1941), the specimen is a rod or stout wire; to this is attached one junction of each of some twenty thermocouples, the other junction being close to, but thermally insulated from, the rod; each couple is in series with its own primary on a mu-metal spiral core, and the low-resistance secondary of many turns is in series with a specially designed galvanometer. The overall sensitivity obtained by the different workers is much the same, ranging from about 200 to 900 erg.cm^{-3} per scale division (100 erg.cm^{-3} corresponds, for iron, to $2{\cdot}8 \times 10^{-6}$ deg.), and it

seems not improbable that the consistency and reliability of the results obtained depend much more on meticulous attention to retails in avoiding or eliminating 'spurious' effects than on the more obvious features of the different methods. An objective test of the reliability of the results is provided by a comparison of the hysteresis loss from the measured heat developed in traversing a cycle, and from the area of the hysteresis loop. In the work of Bates, and of Hardy and Quimby, the agreement is remarkably close, usually within a few per cent.

Examples of the experimental results are shown in the curves labelled Q' in Figures 4.5 and 4.6. These curves give the heat developed in traversing the ascending branch of a hysteresis loop. The aggregate heat developed is half the hysteresis loss. It is to be noticed that the Q' curve for iron (Figure 4.5) is completely different from that for cobalt (Figure 4.6). There is a remarkable diversity in the course of the temperature changes for different materials, as shown in the many curves given by Bates and his collaborators. (If the curves given by different investigators are examined, an equally remarkable diversity is sometimes apparent for effectively the same material, which may suggest that little significance can be attributed to any of these results; but it may be said at once, without implying that all, or even most, of the earlier results are unreliable, that there is no doubt of the substantial accuracy of the results obtained in the work of Bates and of Hardy and Quimby.) Although consideration has been given to the interpretation of the experimental results in some of the papers referred to (e.g. Bates and Weston 1941) and elsewhere (BD. pp. 267–70), the first detailed quantitative theoretical treatment of the analysis of the magneto-caloric data is that recently given by Stoner and Rhodes (1949).

Four interrelated questions have been discussed. The first is how far the 'magneto-caloric properties' of a material may be coordinated with other properties, known or determinable from experiment, involving essentially a thermodynamic treatment. The second is that of the magneto-thermal characteristics of the various elementary magnetization processes, boundary movement, rotation, and change of intrinsic magnetization, which requires consideration of the 'atomic' character of the processes. The third is the application of the general principles to the analysis of the experimental results, that is the determination, as far as may be possible, of just what sequence of elementary processes underlies the observed sequence of temperature changes. The fourth is the question of whether, or to what extent, new or supplementary information might be obtained about the processes occurring as the magnetization changes, which could not be obtained, or not obtained so readily, in other ways. It is possible here to indicate only a few particular points in the treatment, more particularly some of those immediately relevant to the third of the above questions. It may perhaps first be mentioned that the thermodynamics of magnetization often appears extremely involved when it is approached via the relations for a completely general electromagnetic system (Guggenheim 1936, Livens 1945–48); the very simple treatment which is possible when the 'system of interest' is particular magnetizable material in an external magnetic field (Stoner 1935, 1937) is, however, all that is required in the present connection.

The basic equation defining an internal energy function, E, referred to unit volume, of a body in a magnetic field may be taken as

$$dE = dQ + H\,dI, \qquad \ldots\ldots(4.35)$$

where dQ is the heat transferred to the body. For an adiabatic change dQ is zero. A change in internal energy, ΔE, may be divided into two parts, one associated with change of temperature, ΔE_T, measured directly by the heat developed, $\Delta Q'$, given by (4.34), and the other with change of magnetization, ΔE_M. The 'magnetic' energy change is then given by

$$\Delta E_M = \Delta E - \Delta E_T = \int H\,dI - \Delta Q', \qquad \ldots\ldots(4.36)$$

a relation which has been widely used in the presentation of the experimental results. The equations (4.35) and, for adiabatic changes, (4.36) apply whether or not the changes are reversible; they involve only the first law of thermodynamics. Ift he changes are reversible, the internal energy function satisfies the relation

$$dE = T\,dS + H\,dI, \qquad \ldots\ldots(4.37)$$

where dS is the change of entropy. Other thermodynamic functions can then be obtained in the usual way, and among the numerous thermodynamic relations which may be derived, the most important, in the present connection, is

$$\left(\frac{\partial T}{\partial H}\right)_S = -\frac{T}{\rho C_H}\left(\frac{\partial I}{\partial T}\right)_H, \qquad \ldots\ldots(4.38)$$

or, using (4.34),

$$dQ' = -T(\partial I/\partial T)_H\,dH. \qquad \ldots\ldots(4.39)$$

By these equations the change in temperature (or in heat developed) with change of field under isentropic conditions is related to the change of magnetization with change of temperature at constant field. The equations may be applied not only in correlating the immediate observations, where only reversible changes are involved, but also in dealing with the magneto-thermal characteristics of particular types of elementary magnetization process.

Considering only the effects associated with changes in intrinsic magnetization, and changes due to rotations of the magnetization vectors, it is found, by a somewhat lengthy argument, that the heat developed is given by

$$\Delta Q' = a\int d(IH) + b\int H\,dI, \qquad \ldots\ldots(4.40)$$

where

$$a = -\frac{T}{I_0}\frac{dI_0}{dT}, \quad b = \frac{T}{k}\frac{dk}{dT},$$

I_0 being the spontaneous magnetization (or, sufficiently closely, the saturation magnetization) and k the modulus of the effective anisotropy coefficient. Since dI_0/dT and dk/dT are, in general, both negative, a and b are of opposite sign. Increase of intrinsic magnetization is accompanied by a positive contribution to the temperature change on adiabatic magnetization, increase of magnetization due to rotation of the magnetization vectors against the anisotropy field by a negative contribution. An 'explanation' of the first effect, which is satisfactory in principle if not in quantitative detail, can be given in terms of the atomic theory of intrinsic magnetization (I, pp. 65–78); any such explanation of the second effect is at present little more than speculative owing to the incomplete state of development of the atomic theory of anisotropy (§ 2 (iv)). The first term in (4.40) is a generally valid expression for the heat developed owing to changes in intrinsic

magnetization, and the coefficient a is known with reasonable accuracy at least for the ferromagnetic elements. The second term is valid only over reversible ranges of magnetization. If the material consists of an aggregate of effectively unstrained crystalline grains, the value of b can be estimated to the degree of accuracy with which the magneto-crystalline anisotropy coefficients and their temperature coefficients are known, but from the account given earlier (§ 2 (ii)) it is clear that even in the most favourable cases the ideal magnitude of b is much more uncertain than that of a. If the crystal grains are strained, the values of

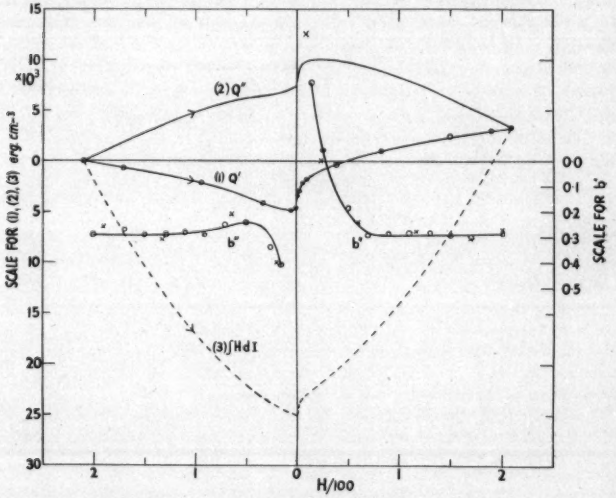


Figure 4.5. Magneto-caloric effect in annealed Armco iron.
Data for curves (1) and (3) from Hardy and Quimby (1938).

Q', heat developed. ($\Delta T \simeq 2{\cdot}8 \times 10^{-4} \Delta Q'$.)
Q'', heat developed less that due to change in intrinsic magnetization.
b'', $\Delta Q''/\Delta \int H\, dI$. Circles, values derived from smoothed curves; crosses, values from experimental points. (Stoner and Rhodes 1949.)

(Reproduced, by kind permission, from the *Philosophical Magazine*.)

k and b are modified, but it would be difficult to determine the magnitude of the modification for, say, a given average value of a tensile or compressive stress, even if the necessary basic data were available. For nickel, for which measurements on the temperature variation of the saturation magnetostriction have been made (Döring 1936), a very rough calculation indicated that for a stress of 1 kg.mm^{-2} there would be an increase in the effective value of k and a decrease in that of b by about a quarter. The relative effect would be smaller for iron and cobalt, for which the magneto-crystalline anisotropy coefficients are larger.

It is now appropriate, before considering other magnetization processes, to turn to the experimental results. A convenient procedure is to express the experimentally observed heat developed in the form

$$Q' = a\int d(IH) + \int b'' H\, dI, \qquad \ldots\ldots(4.41)$$

where the Δ is dropped in the interest of economy of symbols, the first term on the right is the same as that in (4.40), and b'' is to be regarded as a possibly variable coefficient to be determined from the experimental results. Denoting the heat

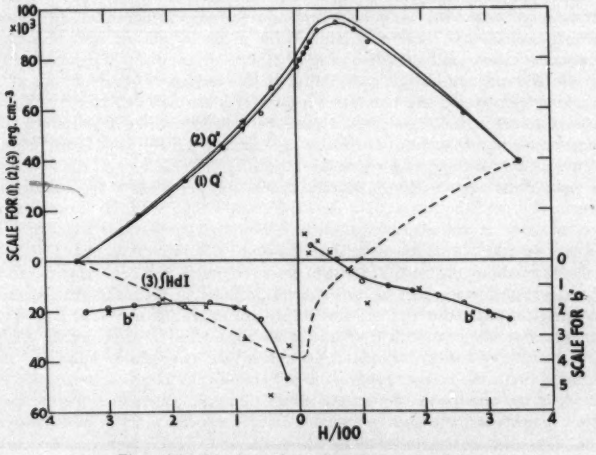


Figure 4.6. Magneto-caloric effect in annealed cobalt.
Data for curves (1) and (3) from Bates and Edmondson (1947).
For symbols, see Figure 4.5. ($\Delta T \simeq 2{\cdot}6 \times 10^{-8} \Delta Q'$.) (Stoner and Rhodes 1949.)
(Reproduced, by kind permission, from the *Philosophical Magazine*.)

developed additional to the reasonably accurately calculable part due to change in intrinsic magnetization by Q'',

$$Q'' = Q' - a\int d(IH) = \int b'' H\, dI. \qquad \ldots\ldots(4.42)$$

Curves showing Q'', and b'' as obtained from (4.42) are given in Figures (4.5) and (4.6) for iron and cobalt respectively. A perhaps unexpected point which is brought out by the calculation of the coefficients a and b in (4.40) is that, although the change in intrinsic magnetization in low and moderate fields may be negligible in its contribution to the change in apparent magnetization, the associated contribution to the magneto-caloric temperature change may be relatively large. The contribution to the heat developed is given by the difference between the Q' and Q'' curves, and it is apparent from Figure 4.5 that for annealed iron the absolute value of this contribution exceeds the absolute value of the aggregate

effect. In cobalt, in contrast, the intrinsic magnetization contribution is relatively small (dI_0/dT is some five times smaller than for iron). When allowance is made for the intrinsic magnetization effect, the residual Q'' curve bears a similar quasi-symmetrical relation to the $\int H\,dI$ curve in both cases.

For iron, the values of b'' for the larger fields are fairly constant at about $-0{\cdot}3$. This agrees as to order of magnitude with the rather uncertain estimated value, from (4.40), for the rotational effect in completely strain-free polycrystalline material of about $-0{\cdot}6$. For cobalt the numerically much higher value of $-2{\cdot}0$ for b'' (Figure 4.6) again agrees as to order of magnitude with the estimated value of $-1{\cdot}3$ for b. It is the agreement as to order of magnitude between the observed and calculated values which should be stressed at this stage rather than the numerical discrepancies, though these may later prove to be significant. At present no definite conclusion can be reached as to whether they are due wholly or in part to errors in the magneto-caloric measurements themselves, to shortcomings in the basic data from which the theoretical values of b are derived, to the effects of residual stresses in the annealed materials, or to the occurrence of processes other than rotations and changes in intrinsic magnetization even in the higher fields.

In the lower field regions there is no question of constancy of the b'' values. The form of the curve, with the marked decrease as zero field is approached, the discontinuity, and the subsequent decrease from a high value to approximately the same value as initially, brings out in a very striking way the effect of those processes not taken into account in the derivation of (4.40). It is, in fact, possible to account qualitatively for the course of the curve in the low field region in terms of the irreversible processes which are there superposed on the reversible processes. In the original paper the magneto-thermal characteristics of irreversible (and also reversible) boundary movement processes are considered in some detail. As far as the analysis of the experimental Q' curves are concerned, however, it is concluded that little further progress is possible until a clear-cut separation can be made of irreversible and reversible effects along the course of a magnetization curve. To this end, the most practicable method appears to be by measurements of the reversible susceptibility along the initial and cyclic magnetization curves over a range of temperatures. From the temperature variation of reversible susceptibility, conclusions might be drawn, using the appropriate thermodynamic relations, as to the magneto-caloric temperature changes associated with the reversible processes, and by subtracting these from the total observed changes, the course of the contributions from irreversible processes could in principle be determined. In recent years comparatively few detailed experimental studies of reversible susceptibility have been made, the most extensive being those of Kirkham (1937) and Kahan (1938). The usual methods of measurements are, however, extremely tedious, and it is hoped that a new rapid oscillographic method developed by Tebble and Corner (1950) will greatly facilitate studies of this kind. It is, of course, essential, particularly in dealing with low field properties, that specimens as nearly as possible of magnetically identical material should be used in related investigations.

With regard to irreversible processes it is perhaps worth while to state that the mere fact of their being irreversible does not necessarily mean that no information can be derived theoretically about their thermal characteristics. If the mechanism of the process is known (the relevant information might be conveyed by curves similar to that in Figure 4.1 (*a*)) the thermal effects can in

principle be calculated. Thus, precise calculations have been made of the energy changes and thermal effects associated with the sequence of reversible and irreversible processes along the theoretical magnetization curve of Figure 4.4 (Rhodes 1948). In general, for actual ferromagnetics, far too little is known about the mechanism or the sequence of the irreversible processes for any theoretical calculation of the associated thermal changes to be possible.

In summary, the reversible temperature changes accompanying adiabatic magnetization, which usually play a larger part in determining the general overall form of the experimental curves (the Q' curves) than the irreversible changes, can be accounted for, in the main, as arising from a superposition of two effects, of opposite sign, one associated with rotational changes, the other with changes in intrinsic magnetization. In cobalt (Figure 4.6) the rotational effect predominates; in some of the alloys studied by Bates and Harrison (1948 a), such as monel metal, with a Curie point not greatly above room temperature, the relatively very large temperature changes, opposite in sign to those in cobalt, are almost entirely due to the intrinsic magnetization effect. In iron and nickel, the magnitudes of the two effects are comparable, and the aggregate effect is small; moreover, since it corresponds to a difference between two numerically larger quantities, relatively large differences may be found in the curves for particular materials subjected to not widely different thermal and mechanical treatments. Although the general features of these curves are now better understood, it cannot be said that any of the curves have been accounted for in quantitative detail. There are two main requirements for further progress in the analysis of these curves: the first is the unambiguous separation of the reversible and irreversible effects, the second the accurate determination of the basic magnetic characteristics of the materials under investigation. If these two requirements could be more nearly met, far more information could be derived from the study of magneto-caloric effects than is at present possible, not only about rotational changes and changes in intrinsic magnetization, but also about boundary movements, and possibly even about boundary formation. The successful surmounting, in recent experimental work, of the many difficulties in the way of obtaining reliable measurements of the very small temperature changes in low and moderate fields is a considerable achievement. It has opened up a new and potentially most valuable method of attack on the numerous unsolved problems which are encountered whenever an attempt is made to give a detailed analysis of magnetization curves.

§ 5. MAGNETIC MATERIALS

5 (i). *Introductory*

It is obviously impracticable in a report of this kind to present in any detail the information which has been obtained in recent years in investigations of the low- and moderate-field characteristics of innumerable ferromagnetic materials. Moreover, a partial compilation of data in tabular form, such as might have been included, would serve little useful purpose, when full compilations of data on magnetic materials generally, and on particular types of material, are readily available. A very large proportion of the time and effort given to research in the general field with which this report is concerned is, however, occupied with the study of the magnetization curves of particular materials. Further, the value of the more fundamental experimental and theoretical investigations discussed in

preceding sections lies, in large part, in their relevance to the whole range of ferromagnetic materials. It seems desirable, therefore, to draw attention in the text, as is done below, to some of the books and articles in which fuller information is collected together, and also at least to 'introduce' some of these interesting materials and the problems which arise in connection with them. Accordingly, after a survey of the general field, with particular reference to the interrelation between what may be called the scientific and technological approaches (§ 5 (ii)), a hurried glance is taken at some of the more recent work on a few illustrative materials falling in the categories loosely described as magnetically soft (§ 5 (iii)) and magnetically hard (§ 5 (iv)).

The 'primary' information about magnetic materials is widely scattered in scientific and technological journals, in technical reports (usually restricted in circulation, and often, in varying degrees, 'confidential'), and in patent specifications. The secondary sources, in the form of books, review articles, and reports of conferences are, therefore, almost indispensable as guides; it may be remarked, too, that the term 'secondary', as applied to them, often does less than justice in its implications. In seeking for information about particular materials, almost invariably, except of course where very recent papers are required, the approach may most expeditiously be made via one or other of these secondary sources. The books and articles mentioned below are included in the references at the end of this report.

Books dealing primarily with ferromagnetic materials are those of Messkin and Kussmann (1932) on the ferromagnetic alloys, the monumental compilation on iron and its alloys by v. Auwers (1938) in *Gmelins Handbuch*, and the short but most informative book by Brailsford (1948). Mention may also be made of a forthcoming book by Hoselitz (1950). The book by Snoek (1947) on recent work in Holland on ferromagnetic materials is naturally much narrower in its scope, but it is particularly useful for the account given of the ferrites. Bitter's *Introduction to Ferromagnetism* (1937) contains a long chapter contributed by Yensen on magnetic materials and their preparation, and Becker and Döring's *Ferromagnetismus* (1939) includes a substantial section (pp. 382–435) on technical magnetic materials and their application. Most of the more general books on magnetism (e.g. Stoner 1934, 1948) have sections on ferromagnetic materials.

Among the more comprehensive review articles dealing with ferromagnetic materials generally are those of Kussmann (1935, 1939) and Webb (1938). Soft materials are surveyed by Yensen (1939, 1948), and hard by Hoselitz (1946). A recent most valuable review of progress on magnetic materials by Brailsford *et al.* (1948) contains sections on electrical sheet steels (Brailsford 1948), permanent-magnet materials (Oliver and Hadfield 1948) and magnetic powder cores (Polgreen 1948). Finally, a very clear and concise account of magnetic materials is given in an Encyclopaedia article by Bozorth (1947). All the above articles include compilations of data in tabular form, diagrams of magnetization curves and associated characteristics of the materials, and, except the last, extensive references.

§ 5 (ii). *General Survey*

Technological requirements. This survey may appropriately be opened by a few remarks on points connected with the technological application of ferromagnetic materials. Most of the work on the preparation of these materials, and on the measurement of their magnetic properties, unlike the corresponding work on

dia- and paramagnetics, has been stimulated, directly or indirectly, by technological requirements. Even a slight acquaintance with the technological background leads to some understanding of the unevenness of distribution over the range of ferromagnetic materials of the vast amount of information which has been accumulated, of the diversity in scientific quality of this information, and of the scientifically unrevealing form in which it is not infrequently presented; at the same time it may lead to a greatly enhanced appreciation of achievements which might be entirely overlooked in a too purely scientific approach to work in this field.

Varied in application though magnetic materials are, their immediate purpose, whether in heavy electrical machines, in transformers large or small, in communication and in general electronic equipment, in electromagnets or in permanent magnets, is to produce magnetic flux more conveniently and more economically than is possible with currents alone. The economic factors are complex, and it is sufficient to note here, with regard to initial costs, that iron is much cheaper than nickel, and nickel than cobalt, so that, other things being approximately equal, preference would be given to materials containing iron as the main constituent. With regard to running costs (leaving aside permanent magnets, whose running costs would ideally be zero), these are less, in 'copper losses', the smaller the applied field necessary to produce the required induction, i.e. the higher the permeability over the relevant range; and when changing, and in particular alternating, fields are involved, they are less, in 'iron losses', the smaller the magnetic hysteresis and eddy current heating.

An upper limit to the induction in the material, and roughly to the external field, due to the ferromagnetic material itself is $4\pi I_0$ (the maximum value of the 'ferric induction') which has values of about 21,600, 6,100 and 17,840 for iron, nickel and cobalt respectively. (It may be noted in passing that there is little purpose in using an iron core in electromagnets for fields much larger than 25,000 oersteds, say for fields of the order of 10^5 oersted, for most of the field must be produced directly by currents. Air core coils are used with heavy currents of short duration (Kapitza) or, for continuous fields, with elaborate cooling arrangements (Bitter).) A fair relative measure of the potential range of application of a ferromagnetic material is therefore given by its saturation magnetization, I_0. The values of I_0, as fundamental characterizing magnitudes of ferromagnetics, have already been fully discussed (in I) in relation to elements and alloys. The value of I_0 for iron is appreciably exceeded only in some of the alloys in the iron–cobalt system (cf. I, p. 92); it is the combination of high saturation intensity and very low cost which gives to iron its outstanding technological importance as the basic ferromagnetic material.

Within the limits imposed in the choice of materials by the maximum induction required, the general desiderata (except in permanent magnet materials) are low hysteresis and, over the relevant range of field or induction, high permeability. This report has largely been concerned with the fundamental factors on which these depend. The characteristic magnitudes which are important are the crystal anisotropy (directly, and also indirectly in connection with the effects of inclusions) and the magnetostriction (mainly indirectly in connection with the effects of internal stresses). Alloys give a much wider scope than the elements for obtaining suitable combinations of characteristics, mechanical and electrical as well as magnetic. For many applications their often much higher resistance than that of

pure metals is an immediate advantage in reducing eddy current losses. The use of materials in special forms or states of aggregation (e.g. laminations, 'dust' cores) should not be overlooked as a means of obtaining effective electrical and magnetic characteristics which more nearly meet particular requirements. (For a recent review of the applications of powder metallurgy to magnetic materials see Pfeil *et al.* 1947.)

Pure iron. The sensitivity of the low-field properties of a ferromagnetic to small disturbing influences is well brought out by the work of Yensen, extending over many years, Cioffi and others on the effect of further purification (by high temperature treatment *in vacuo*, or in a suitable atmosphere) on already pure iron, say 99·9 per cent Fe. (For general references, see § 5 (i).)

Studies on very pure single crystals have been made by Cioffi, Williams and Bozorth (1937). The initial material, an almost spectroscopically pure sample of carbonyl iron, was heat treated in hydrogen at 1,500° C., cooled to 880° C. and maintained at that temperature to permit the growth of large crystals. From the sheet a picture frame crystal (cf. § 2 (ii)), approximately 1 cm. square in outside dimensions, was cut, with sides parallel to a [100] direction. Magnetization curves were obtained with this specimen after further annealing. The maximum measured permeability was $1{\cdot}43 \times 10^6$ at $B = 17{,}500$, corresponding to the attainment of 80 per cent of the saturation intensity in a field of 0·012 oersted. This is the highest permeability so far recorded. Surprisingly, the measured coercivity, $H_c = 0{\cdot}15$, was considerably higher than values which have been obtained for less pure polycrystalline iron (cf. Bozorth 1937).

For polycrystalline iron, a maximum permeability of $2{\cdot}8 \times 10^5$ is given for hydrogen-annealed Armco iron by Cioffi (1934), but the highest values obtained in the many very pure specimens prepared and investigated by Adcock and Bristow (1935) and by Cleaves and Hiegel (1942) are about 3×10^4 and 9×10^4 respectively. These values are all considerably higher than the normal value of about 6×10^3 for dynamo sheet. There are many sources of uncertainty in the determination of high permeabilities in very low fields. Moreover, although measurements of maximum permeability, μ_{max}, may be somewhat less uncertain than those of the initial permeability, μ_0, the quantity itself depends on so many factors that it has no clear-cut theoretical significance. The values that have been quoted serve to show that very high permeabilities are not a prerogative of special alloys, but can occur equally in a pure, strain-free ferromagnetic element.

High initial and maximum permeabilities and low coercivity are almost invariably associated, and the values of any or all of these give a measure of what is qualitatively referred to as the magnetic 'softness' of the material. Pure iron itself, quite apart from its costliness in production, would be of very limited technological value as a magnetic material, for the magnetic softness is associated with mechanical softness and low electrical resistance, and also with defects of a chemical character. Alloying does not improve the magnetic softness; but in alloys comparable magnetic properties may be obtained much more readily than by intensive purification, and in combination with other required characteristics.

Alloys. Alloys are broadly classified as homogeneous (single-phase) or heterogeneous. The classification is generally convenient, but difficulties arise in connection with 'traces' of impurities. On practical grounds, Yensen (1937, p. 72) prefers to place in the heterogeneous class only alloys containing microscopically visible inclusions (of volume greater, say, than 10^{-16} cm^3). From a

magnetic standpoint such a classification would be misleading, for the magnetic effect of sub-microscopic segregates may be enormously different from that of the same number of atoms uniformly distributed in strict solid solution. Leaving aside the various special problems which arise in this way, it may be said that over a single-phase range starting from a ferromagnetic element the variation of magnetic properties proceeds smoothly not only in respect of such basic characteristics as the saturation magnetization (cf. I), the crystal anisotropy (cf. §2(ii)) and magnetostriction, but also in respect of the softness characteristics, μ_0, μ_{max} and H_c. Transition to a heterogeneous state (e.g. in passing a solubility limit) is in general marked by a rapid change in these latter characteristics. Little can be said that is both concise and general, however, about the low field properties of heterogeneous alloys, for these properties depend so much on the character of the phase segregation, and hence on the details of the thermal treatment to which the alloys are subjected. Each alloy raises its own complex of problems. There is little doubt that, with the advance of magnetic theory, magnetic investigations will contribute more and more towards the solution of problems which are regarded as primarily metallurgical. This is well illustrated by the work of Gerlach (1938) on the effect of different thermal treatments on the coercivity of dispersion-hardening alloys (Ni–Au, Ni–Be, Fe–Cu), and by that of Sucksmith (1939) on the saturation magnetization and its temperature variation in alloys in the iron–nickel–aluminium and other systems (cf. I, pp. 106–9) and more recently on the low field properties of a series of heterogeneous alloys (see §5(iv)).

Among the soft magnetic materials by far the most important technologically are the iron–silicon alloys and the nickel–iron alloys (see §5(iii)). The former alloys came into use soon after the beginning of the century, and the latter in the early twenties. The earlier, and very largely the later, 'discoveries' of alloys with good magnetic qualities are in the main due to systematic empirical investigations. It is only in comparatively recent years that 'theory', in providing at least partial explanations of known properties of particular alloys, has, to a corresponding extent, served as a guide to methods of 'improving' existing alloys, and, to a much smaller extent (since at least the economically favoured range of alloys has been well covered empirically), to the discovery of new ones. A scientific approach to this problem is well illustrated by some work of Snoek and Went (Snoek 1947, pp. 14–20). Magnetic softness is favoured by low magnetostriction and low magneto-crystalline anisotropy. In binary alloy systems there may be zero-λ and zero-K compositions, but it is unlikely that K and λ will be zero at the same composition. In ternary alloys the zero points become zero lines. An investigation of the course of these lines may indicate whether or not alloys of outstanding qualities in softness are likely in the ternary range. Snoek and Went investigated (or supplemented previous investigations on) the variation of λ and K with composition of eleven binary iron alloys in the single-phase body-centred range. A zero-λ composition was found only in the Fe–Si system (at about 6 per cent Si by weight) and zero-K compositions only in the alloys with Si, Al, Co. This suggests that coincidence of zero λ and zero K might occur in the ternary systems Fe–Si–Al and Fe–Si–Co. The course of the zero lines in these systems could be determined sufficiently closely by examination of a fairly small number of ternary alloys. The lines do not intersect in the Fe–Si–Co diagram, so that no alloys with unusually high permeability are to be expected in this ternary system. In the Fe–Si–Al diagram the curves intersect roughly at

the composition Fe 84, Si 10, Al 6. It happens that the high permeability (μ_{max} up to $1{\cdot}6 \times 10^5$) and low hysteresis loss of alloys in this composition region had already been found in a most extensive investigation of this ternary system by Masumoto (1936) (the alloys, which, under the name of 'sendust', are used in powder cores, have been further investigated by Selisski and Zaimovski (1941)), so in one sense nothing new had been discovered; the newness is in the method of approach, which is of a kind which may well be more widely applied in the future. The importance of the λ, K factors will be further exemplified below (§ 5 (iii)). Other factors than these are, however, involved; in the iron–silicon alloys neither λ nor K is unusually small.

Heterogeneity seems to be a necessary condition for magnetic hardness (essentially, high coercivity, though other characteristics are also required for permanent-magnet materials). In the martensitic carbon steels (almost universally used for permanent magnets prior to 1910), and the tungsten and chromium magnet steels, there is little doubt that inhomogeneity of internal stress is the major factor determinative of the coercivity (H_c values up to about 80) through its direct or indirect effect on the positional variation of domain boundary energy (§ 4 (ii)). In the many alloys developed from the iron–nickel–aluminium prototype since 1931, with much higher coercivities ($H_c > 500$), stress and inclusion effects may be involved, but it seems probable that the single-domain mechanism (§ 4 (iv)) is a more important factor. Some of these alloys are considered below (§ 5 (iv)) and also the 'micro-powder' magnet materials (which, it may be noted, are in one sense heterogeneous alloys with air, or a non-ferromagnetic binding material, as the second phase). The commercial magnet alloys are usually very complex in composition, and for a scientific investigation of the underlying processes taking place other alloys may be much more suitable. Examples are the binary Ni–Au, Ni–Be and Fe–Cu alloys already mentioned, and Fe–Mo alloys studied by Sucksmith and Margerison (1949), in which there is a one-phase state at high temperatures, which may be retained by quenching, and a well defined equilibrium two-phase state at ordinary temperatures, so that the extent and character of the precipitation may be controlled by appropriate heat treatment.

Many ferromagnetic binary and ternary alloys are now known which contain no ferromagnetic element. The saturation magnetization, I_0, is seldom higher than that of nickel, but some of the alloys, after appropriate heat treatment, have striking low field characteristics. Thus the Heusler type alloy 'silmanal' (Ag 86, Mn 9, Al 5) has a remanence, B_r, of 550, and a coercivity, ${}_IH_c$, of 6000 (${}_BH_c = 540$). The main interest of these alloys, however, is in their intrinsic magnetization characteristics, and since the problem of their low field properties is similar in kind to that of other more fully investigated alloys, they will not be considered here. (For references, see Bitter (1937), Brailsford *et al.* (1948), Sucksmith (1949 a).)

Non-metallic ferromagnetics. Although far less work has been carried out on the non-metallic than on the metallic ferromagnetics much of it is of such a character that an account of it sufficiently detailed to be useful could not conveniently be incorporated in the present or the earlier report (cf. I, p. 109). Mention must be made, however, in a section on magnetic materials, of the intensive investigation of the ferrites carried out by Snoek, Boeke, Louwerse and others largely during the period 1940–45, and described by Snoek (1947, pp. 20–24, 68–99). The ferrites have the general formula $MO.Fe_2O_3$, and those

of cubic structure which were primarily studied, with Mg, Zn, Cu, Fe, Co and Mn for M, form mixed crystals in any proportion. For most of these ferrites the saturation intensity is of the same order as that for nickel. Zinc ferrite is exceptional in that it is not ferromagnetic at room temperature, the Curie temperature being lower. The technological stimulus to this work is due to these materials having a very high specific resistance (of the order 2×10^4 ohm cm.) compared with metals, so that in this respect they have very favourable properties for core materials, particularly for high frequencies. The initial permeability of the simple ferrites, as ordinarily prepared is, however, rather low (e.g. less than 100 for a carefully prepared sample of magnetite, FeO. Fe_2O_3, and usually less than 10 for an ordinary sample). It was found that greatly increased values of μ_0 (and associated low values of H_c and of hysteresis loss) could be obtained in mixed crystals with zinc ferrite as one of the components. With the non-ferromagnetic component, the Curie point in the mixed crystal can be reduced to a temperature not far above room temperature, so that the usual maximum in the (μ_0, T) curve is brought into the room temperature region. At the same time the effective anisotropy and magnetostriction are reduced to zero or to very small values. Values of μ_0 up to 4,000 have been obtained in the nickel–zinc ferrites, which may be compared with typical values of a little over 100 for nickel–iron alloy powder material, and of well under 100 for most iron powders. Detailed investigations were made of the effects of different heat treatments and of the proportions of simple ferrites required to give optimum properties for various purposes. In the preparation of homogeneous mixed crystals many difficulties were encountered, which were overcome by the development of carefully controlled sintering processes. For a full account of many interesting details of this work, in which science and technology are so happily blended, and which has led to the production of various types of 'ferroxcube', already widely applied, particularly in high frequency work, reference may be made to Snoek's book.

5 (iii). *Iron–Silicon and Nickel–Iron Alloys*

A general understanding of the magnetic characteristics of the iron–silicon and nickel–iron alloys may perhaps be facilitated by a brief qualitative restatement of the main factors which, in the light of both theory and experiment, appear to be determinative of magnetic softness. This depends essentially on the freedom of movement of domain boundaries. The freedom is checked (as fully discussed in § 4 (ii)) by inhomogeneities of composition and of internal strain. Internal strains of mechanical origin can be minimized by appropriate annealing treatments. Beyond this, more and more complete removal of impurities enhances the permeability and decreases the coercivity, as exemplified in the work on very pure iron (§ 5 (ii)). The effect of variations in composition (in particular those coraesponding to the presence of inclusions, either non-ferromagnetic or less ferromagnetic than the matrix) is less the smaller the magneto-crystalline anisotropy; the effect of internal stresses (whether arising from lattice distortion due to the presence of impurities, or as a result of thermal and mechanical treatment) is less the smaller the magnetostriction. Small values of K and λ are thus favourable to magnetic softness, and the actual values set a limit to the values of the permeability (or at least of the initial permeability) that are attainable, for there is always some degree of heterogeneity of stress resulting from magnetostrictive changes as the material cools down from the Curie point (cf. I, p. 65). This limit is usually

very much higher than the values ordinarily obtained; within this limit, the main controlling factor appears to be the amount of impurity present, and, of course, its character, and form of aggregation.

In brief commentaries on various alloy systems detailed references to original papers would be out of proportion, and only a few supplementary references are given. Much more extensive accounts of these systems, with full references, are readily available in the books and papers listed above (5 § (i)), among which attention may be particularly directed to the book and review article by Brailsford (1948).

It is impracticable to include magnetization curves of all the materials mentioned in the present account, but an appreciation of the significance of the numerical values quoted may be assisted by the representative curves for soft magnetic materials shown in Figure 5.1 (based on a figure given by Stoner (1944)).

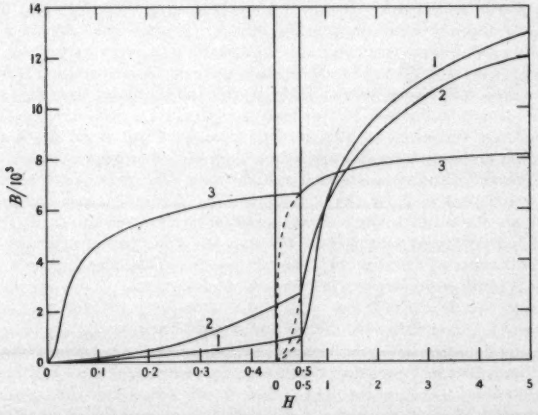


Figure 5.1. Representative magnetization curves for soft magnetic materials.

1. Swedish iron, annealed (C 0·027, Mn 0·030). $4\pi I_s$, 21,200. H_c, 0·8. μ_0, $0{\cdot}47\times10^3$. μ_{max}, $6{\cdot}4\times10^3$.
2. Silicon sheet steel, hot rolled (Si approx. 4·0). $4\pi I_s$, 19,500. H_c (for B_{max} 10,000), 0·5. μ_0, $1{\cdot}25\times10^3$. μ_{max}, $6{\cdot}5\times10^3$.
3. Permalloy C (approx. Ni 78·5, Fe 18·0, Mo 3·0, Mn 0·5). $4\pi I_s$, 8,200. H_c (for B_{max} 5,000), 0·03. μ_0, 2×10^4. μ_{max}, 7×10^4.

Iron–silicon. The solid solution range of silicon in iron extends to about 15 per cent silicon (by weight), but the brittleness of the alloys increases rapidly with silicon content above about 4 per cent. Alloys with above 6 per cent are seldom used, the usual range in electrical sheet steels being up to about 4·5 per cent. The high-temperature face-centred region (or γ-phase, extending in iron from about 910° to 1390° c., cf. I, p. 83) is diminished by addition of silicon; alloys with more than about 2·5 per cent Si are single-phase body-centred over the whole temperature range up to the melting point, and large-grained material can be readily obtained. (For recent work on the phase diagram, see Guggenheimer

et al. 1948). As illustrative of the effect of alloying, for a 4 per cent alloy, I_0 is about 10 per cent smaller than for pure iron, the anisotropy coefficient about 40 per cent smaller (cf. §2(ii)), and the specific resistance about six times as great (about 55 as compared with 9·6 microhm cm.). The magnetostriction does not appear to have been studied for single crystals. The available data indicate that the mean λ passes from positive to negative between 6 and 7 per cent Si. The reduction in eddy current losses due to the addition of silicon is readily intelligible. It does not seem, however, that the increased permeability and lower coercivity (and lower purely magnetic hysteresis loss) as compared with iron containing similar impurities can be accounted for purely in terms of the not greatly lower K and λ values. It is usually attributed to the 'scavenging' action of silicon on oxygen impurities, and to the effect of dissolved silicon on carbon impurities; these tend to precipitate as graphite, which has a less harmful magnetic effect than finely dispersed cementite. The losses may also be reduced to some extent by the increased grain size obtainable. Part of the hysteresis loss is usually attributed to a grain boundary effect, and some of the results obtained suggest (though this is a controversial issue) that in otherwise similar materials, this contribution to the loss is proportional to the grain boundary area. The loss does appear to decrease with increasing grain size, but a limit is set to improvement in this direction by deterioration in the workability of thin sheet material if the linear size of the grains approaches the sheet thickness. The effect of impurities and grain size has been discussed by Richer (1938) who concluded that, with the minimum impurities present in commercial 4 per cent silicon steel, the lower limit to the total loss (estimated at 0·95 watt kg^{-1} at 50 cycles, corresponding to about 1,500 erg cm^{-3} per cycle, for $B_{max} = 10{,}000$) was not far below that obtained in practice (about 1·3 watt kg^{-1}). The type of commercial silicon steel referred to here is that produced by a hot-rolling process, and of which the qualities, according to Brailsford (1948, p. 528), 'have remained unchanged over a long period of years'.

Sheet steel of considerably better quality has been successfully produced commercially (under the name of 'Hipersil') by a combination of two methods. The first is essentially a reduction of impurities, particularly of carbon, by a high temperature anneal in pure dry hydrogen. This operation is preceded by an appropriate sequence of cold-rolling and annealing processes, by which sheets are obtained with a preferential orientation of the crystal grains, approaching that in which the [100] directions are in the rolling directions, and the [110] directions at right angles in the plane of the sheet. The sheet is anisotropic with an easy direction along the rolling direction (the magnetization curves are similar to, but less sharp than, those for single crystals of iron, as shown in Figure 2.1); for magnetization in this direction, the induction is higher for a given field than for isotropic sheets, and the losses lower. The improvement is considerable. Thus the total iron loss for cold-rolled sheet containing about 3 per cent Si, under the conditions given above, is about 0·6 of the loss for the corresponding hot-rolled isotropic sheet. In unfavourable directions the losses may be greater than for isotropic sheets, but for transformers, the main application, almost full advantage may be taken of the greatly improved qualities along the rolling direction. In the country of its origin cold-rolled sheet is rapidly replacing hot-rolled, so it would seem that the greater initial cost is more than compensated by the lower running cost of the finished transformers.

The mechanism by which cold rolling gives rise to preferential orientation in subsequent recrystallization is still very obscure, but such effects are not, of course, peculiar to ferromagnetic alloys (cf. Barrett 1943, pp. 395–434). That preferential orientation should give rise to improved magnetic qualities in the sheet may seem, at first, readily intelligible. Quantitatively, however, there are many unsolved problems, perhaps more in connection with eddy current losses than with the actual shapes of the magnetization loops, and investigations on these questions are in active progress (see, for example, Lawton, Shoenberg and Stewart 1947, De Barr 1949). The orientational distribution of crystals in sheets may be determined by x-ray, optical and other methods. In certain cases the degree of preferred orientation can be rapidly estimated by the torque curve method (cf. §2(ii)). The possibility of estimating the distribution from the magnetization curves themselves has been explored by Richer (1944, 1947). Such a method, suitablydeveloped, might give valuable supplementary information, but the 'unknowns' at present are far too numerous (as pointed out by De Barr in 1948, unpublished) for an unambiguous determination to be made by this method alone.

Nickel–iron alloys. A general survey of the magnetic properties of the nickel–iron alloys and their applications, with extensive data in tabular and graphical form, has recently been published by the Mond Nickel Company (see Anon., 1949). This forms an admirable introduction to the more detailed discussions of particular questions in the books and review articles already mentioned, as well as to the accounts of original work. The phase diagram of the binary alloys has already been discussed (I, pp. 104–6). The alloys to be considered here are exclusively those in the single-phase face-centred range extending from about 35 per cent to 100 per cent nickel. The saturation intensity, I_0, starting from nickel, increases almost linearly from about 485 for Ni to about 1230 for 50 Ni 50 Fe, and subsequently decreases rapidly. These alloys do not compete with iron-rich alloys in transformers and power equipment owing to their much lower saturation magnetization (for iron, $I_0 \simeq 1720$), but where high permeability and low losses are required at low and moderate inductions they have found innumerable applications.

The earlier investigations, notably by Arnold and Elmen (1923), showed that very high permeabilities could be obtained in alloys in the range 75–80 Ni by suitable heat treatment. (The binary alloys will be specified by the percentage of nickel by weight unless otherwise stated.) Annealing at about 1000° c. and slow cooling is ineffective, but subsequent heating to 600° c. followed by air-quenching gives high permeability with a fairly well defined maximum at about 78·5 Ni ($\mu_0 \simeq 1{\cdot}2 \times 10^4$, $\mu_{\max} \simeq 9{\cdot}0 \times 10^4$). (It may be noted, to prevent misunderstanding, that although 'permalloy' is strictly a trade name for a series of alloys of this basic composition, the high permeability nickel–iron alloys generally are sometimes referred to as 'the permalloys'.) The saturation magnetostriction, λ, of the iron–nickel alloys passes from positive to negative in this region, and it was suggested by McKeehan (1925) that the high value of μ_0 might be associated with the low value of λ. This suggestion was later made more precise by Kersten's relation for the maximum value of μ_0 (cf. I, p. 65), namely

$$(\mu_0)_{\max} - 1 = 4\pi(\kappa_0)_{\max} = \frac{8\pi}{9}\frac{I_0^2}{\lambda^2 E}, \qquad \ldots\ldots(5.1)$$

where E is Young's modulus. The work of Lichtenberger (1932), however, shows that the mean value of λ and the values along the main crystal directions all pass through zero at approximately 82 Ni. At 78·5 Ni, the magnetostriction is, in fact, comparable with that for iron. The high permeability is not, therefore, to be explained solely in terms of the magnetostrictive characteristics of the material. The magneto-crystalline anisotropy must also be considered. The leading anisotropy coefficient, K_4, passes through zero at about 75 Ni (see Figure 2.5) from the results of Kleis (1936), confirmed by the more detailed work of Grabbe (1940). It may be that, with a minimization of impurity precipitation by the air-quenching treatment, the maximum μ_0 at about 78·5 Ni is to be explained mainly as resulting from an optimum combination of low values of both K and λ.

It has often been suggested (e.g. Becker and Döring 1939, p. 412) that the possible formation of a superlattice at and near the composition Ni_3Fe (75·9 Ni) might have some connection with the high permeability of alloys in this range. Although the character of the dependence of many of the properties on thermal treatment suggests that an ordering process may be involved, the definite establishment of its occurrence by x-ray methods is difficult owing to the similarity in scattering power of iron and nickel. It was first accomplished, using cobalt $K\alpha$ radiation, by Leech and Sykes (1939), who obtained films showing all the superlattice lines with a specimen annealed for 500 hours in the temperature range 490° to 370° c. Among the many investigations of the effect of ordering on magnetic, electrical and thermal properties (e.g. Kaya 1938, Haworth 1938, Grabbe 1940, Källbäck 1948, Sucksmith 1949 b), mention may be made particularly of the work of Grabbe, who, using single crystals, measured the magneto-crystalline anisotropy (cf. § 2 (ii)) and other properties of a series of alloys in the range from 65 to 80 Ni, both in the quenched state and after prolonged annealing. It may be said at once that ordering, or the thermal treatment required to produce ordering, so far from increasing the permeability, decreases it enormously, as would be expected from the earlier work. Grabbe suggests, very plausibly, that the lower permeabilities resulting from slow cooling, as opposed to air-quenching, are due to the heterogeneity arising from small ordered regions in an otherwise disordered lattice. The change in the anisotropy coefficient, K_4, on ordering is greatest (from 0 to $-2{\cdot}3 \times 10^4$ erg. cm^{-3}) at about 75 Ni. The composition at which K_4 is zero in the ordered alloy is between 65 and 68 Ni, as compared with about 75 Ni for the disordered alloy. This, as pointed out by Snoek (1941) and others, may be relevant to the explanation of the effect of magnetic annealing.

It was found by Dillinger and Bozorth (1935) that the maximum permeability of a certain range of alloys in the iron–cobalt–nickel system could be considerably increased by annealing in a magnetic field. The largest effects are produced when the field is applied over the temperature range from 500° to 400° c., and the Curie point must be substantially above the lower of these temperatures, which may be identified with that at which plastic flow begins to occur. In the binary nickel–iron alloys the greatest change is produced for compositions between 65 and 70 Ni, where the Curie temperature rises to a maximum of about 600° c. A typical increase in μ_{max} for a 65 Ni alloy is from 5×10^3 to 3×10^5, and for one specimen, which was given a preliminary prolonged high temperature anneal, the change was from 2×10^4 to 6×10^5. With a single crystal, in hollow rectangle form, of similar material, Cioffi, Williams and Bozorth (1937) obtained a change

in μ_{max} along the [100] direction from $1{\cdot}8 \times 10^4$ to the enormous value of $1{\cdot}33 \times 10^6$, nearly as great as that obtained in a single crystal of pure iron (see also Bozorth 1937). The magnetically annealed alloys are strongly anisotropic. In one particular specimen, μ_{max} along the direction in which the field had been applied was greater by a factor of 150 than that at right angles. The tentative qualitative explanation of these effects suggested by Dillinger and Bozorth is simple. As the spontaneous magnetization increases on cooling below the Curie point, magnetostrictive stresses are set up, which may be relieved by deformation if the cooling over the plastic range is sufficiently slow. The deformation is such as to make the energy a minimum for the actual direction of magnetization over any uniformly magnetized region, whether this is an entire single crystal grain or a smaller domain. When the material has 'hardened' at lower temperatures, these actual directions become relatively easy directions of magnetization compared with others which, in an unstrained crystal, are equivalent. Ordinarily, these easiest directions will be distributed at random among the equivalent easy directions of the crystals constituting the polycrystalline aggregate; but if a field is applied during the cooling, the easiest directions will be those most nearly parallel to the applied field. In the demagnetized state the domain magnetization vectors would then be approximately parallel or antiparallel to the previously applied field, and change of magnetization in this direction could occur readily by the movement of 180° boundaries, giving a high initial and maximum permeability; a transverse change of magnetization would involve a turning of domain magnetization vectors from the easiest to more difficult directions, and the permeability would be correspondingly low. It is maintained by Becker and Döring (1939, pp. 418–22) that this treatment breaks down quantitatively, and they suggest speculatively that a change of crystal structure from cubic to tetragonal may be involved; whether or not this is so, there can be little doubt that the basic process is much as envisaged by Dillinger and Bozorth. It may be noted that a value of λ different from zero is essential for the magnetic annealing effect to occur at all; on the other hand, the smaller the value of K, the greater is the effect likely to be, as the directions of easiest magnetization developed during cooling are more completely controlled by the direction of the applied field. The relevance of Grabbe's discovery of the shift of the zero K region in the ordered nickel–iron alloys at once becomes apparent. It also becomes understandable that only a very small change is produced in the permeability of iron and iron–silicon, with their relatively high K values.

The cold-rolling process, described in connection with the iron–silicon alloys, was first applied with success to the nickel–iron alloys. The details,however, cannot be described here. Short accounts of the work are given by Brailsford, and representative papers are those of Snoek (1935, 1936), Randall (1937) and Scholefield (1949). For accounts of the many interesting ternary alloys, such as those in the Fe–Co–Ni system which include the 'perminvars', with effectively constant permeability over a low field range, reference may be made first to the books and articles already mentioned. A number of the technologically valuable alloys, although they have more than two components, are perhaps best regarded as nickel–iron alloys 'modified' by the replacement of some of the nickel or iron by other elements to obtain better magnetic characteristics for particular purposes, to increase resistivity, to improve workability, or to simplify commercial production by, for example, decreasing the strictness of control necessary in heat treatment.

Examples are the chromium and molybdenum permalloys (78·5 Ni, 17·7 Fe, 3·5 Cr or Mo), mumetal (76 Ni, 17 Fe, 5 Cu, 2 Cr), and '1040' (72 Ni, 11 Fe, 14 Cu, 3 Mo). Finally mention may be made of a recent new material of high permeability, described by Boothby and Bozorth (1947), to which the name 'supermalloy' has been given. The composition is about 79 Ni, 15 Fe, 5 Mo, 0·5 Mn. By a combination of hot and cold rolling, high temperature (1300° c.) annealing in pure dry hydrogen, and cooling at a critical rate over the range 600° to 300°, values of μ_0 from 5×10^4 to $1{\cdot}5\times10^5$, and of μ_{max} from 6×10^5 to $1{\cdot}2\times10^6$ have been obtained. The coercivity ranges from 0·002 to 0·005, and the hysteresis loss per cycle at $B=5{,}000$ is less than 5 erg. cm^{-3}. The remarkable properties are believed to be due on the one hand to a minimization of the usual impurities, and on the other, though this has yet to be confirmed, to a combination of small λ and small K, resulting from a critical degree of ordering controlled by the rate of cooling. The figures given refer not to a laboratory specimen, but to a material which has been made in considerable quantities and used in the form of thin tape for the cores of communication transformers. The development of supermalloy illustrates again how the meeting of technological requirements may be assisted by a scientific approach to the problems involved.

5 (iv). *Permanent-magnet Materials*

Required characteristics.

It is perhaps desirable to introduce permanent-magnet materials by a few remarks on the criteria used in assessing them. The most useful single number criterion of the suitability of materials for permanent magnets is the maximum value of the product BH on the descending branch of the hysteresis loop, as shown in Figure 5.2 (based on a figure given by Stoner (1944)).

By a readily proved general theorem, in the absence of currents,

$$\int BH\,dv=0, \qquad \ldots\ldots(5.2)$$

where the integral is taken over all space (or, approximately, over a volume at the boundary of which the field is approximately zero). For a single magnet,

$$\int H^2\,dv_o=-\int BH\,dv_i, \qquad \ldots\ldots(5.3)$$

where the subscripts o and i refer to the space outside and inside the magnet. A somewhat artificial but often useful distinction may be made between the useful flux (e.g. in a pole gap) and the leakage flux. Denoting the volume of the magnet by V, the effective volume of the gap by v, and the field in the gap by H_g,

$$-\overline{(BH)}\,V=\int H^2\,dv_o=q\overline{H_g^2}v, \qquad \ldots\ldots(5.4)$$

where q is a 'leakage' factor necessarily greater than, and in practice usually much greater than, unity. It will be clear from these equations, first, that the greater the value of $(BH)_{max}$ the greater the potential efficiency of the material for magnets (i.e. the smaller the volume required to produce a given effect) and, secondly, that if full advantage is to be taken of the properties of the material the magnet should be so designed that the mean working point on the demagnetizing curve is at or near the $(BH)_{max}$ point. The mechanical properties of a material and its cost must be taken into account in assessing its overall efficiency; but apart from these, if the shape of the magnet is determined in part by other than

purely magnetic considerations (e.g. that it must fit conveniently into the general assembly design of a piece of equipment), the material of highest $(BH)_{max}$ value which is available is not necessarily the most efficient.

Dependence of $(BH)_{max}$ on remanence and coercivity.

High $(BH)_{max}$ values are ordinarily associated with high values of both the remanence, B_r, and the coercivity, H_c, but it must be borne in mind that the relevant coercivity is the value of H at which B is reduced to zero, ${}_BH_c$, and not the value, which may be much higher, at which I is reduced to zero, ${}_IH_c$. In the

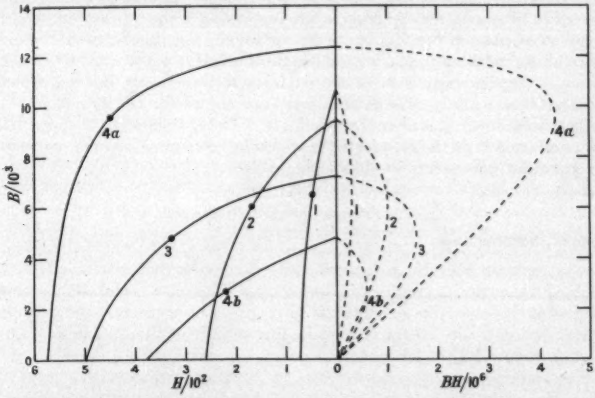


Figure 5.2. Representative demagnetizing and BH curves for permanent-magnet materials.

1. Tungsten steel (W 5·5, C 0·6).
2. Cobalt steel (Co 35, Cr 6, W 4, C 0·8).
3. Alnico (Ni 18, Co 12·5, Al 10, Cu 6).
4. Alcomax (Co 25, Ni 11, Al 7, Cu 6).

Alcomax is anisotropic. The curves show typical characteristics along (4 *a*), and at right angles to (4 *b*), the direction of the field applied during cooling. The $(BH)_{max}$ points are marked on the curves.

remainder of this section, unless otherwise stated, H_c or 'coercivity' refers to the value ${}_BH_c$. The shape of the descending (B, H) curve is important, and is sometimes roughly indicated by a 'fullness factor', γ, defined by

$$\gamma = (BH)_{max}/(B_rH_c). \qquad \ldots\ldots(5.5)$$

Estimates may be made of the highest $(BH)_{max}$ values ideally attainable, following a method of treatment of the geometry of demagnetizing curves developed in detail by Hoselitz (1944). The highest values would occur if the intensity of magnetization, I, remained constant at the remanence value, I_r, as the field increased negatively. The induction would then be

$$B = 4\pi I_r - H = B_r - H,$$

and a limit to $(BH)_{max}$ is given by

$$(BH)_{max} \nearrow \tfrac{1}{4}B_r^2. \qquad \ldots\ldots(5.6)$$

If, however, H_c is less than $\frac{1}{2}B_r$, simple consideration shows that

$$(BH)_{max} \geqslant H_c(B_r - H_c). \qquad \ldots\ldots(5.7)$$

Throughout this section it will be convenient to express $(BH)_{max}$ values in units of 10^6 erg. cm^{-3}. There are now available commercial materials for which the $(BH)_{max}$ values approach 6. It follows at once from (5.6) that such values cannot be exceeded by materials for which B_r is less than about 5,000. Thus for the Heusler type alloy silmanal already mentioned (§5(ii)), the B_r value is about 550, and in spite of the enormous $_IH_c$ value 6,000, the maximum possible value of $(BH)_{max}$ is 0·076, which is practically attained (the observed value is, in fact, given as 0·08). For most of the permanent magnet alloys, H_c is much less than $\frac{1}{2}B_r$, and the limit for $(BH)_{max}$ is that given by (5.7). This limit is approached only by the anisotropic alloys of the alcomax type. Thus for one of the alloys of the alcomax series, with B_r 12,500, H_c 550, the observed value of $(BH)_{max}$ is 4·3 as compared with the limiting value of 6·6, a ratio of 0·65. For isotropic alloys the ratio is smaller. For alnico, for example, the ratio is about 0·48.

Ideal permanent-magnet material.

A rough estimate of an ideal upper limit to $(BH)_{max}$ is obtained by taking B_r as 2×10^4 in (5.6), which gives a value of 100, more than 10 times as great as that for any of the materials available at present. The increase in field for a given ratio of magnet volume to gap volume would be by a factor of about 3, the improvement being comparable with that from tungsten steels to present-day magnets. For such a value to be attained, however, a material would be required with a remanence equal to the saturation magnetization of iron, a rectangular (I, H) loop, and a coercivity of 10^4 oersted. Consideration has been given to a highly idealized magnetic material on the one hand to show that considerable improvement may still be made on materials now available, and on the other to make it clear that even the ideally possible improvement (which greatly exceeds any 'theoretically possible' improvement which has so far been envisaged) is not nearly so great as is sometimes supposed.

The characteristics of actual permanent-magnet materials are so comprehensively tabulated in the articles already mentioned, particularly in that of Oliver and Hadfield (1948), that no more than a bare indication of them is given here, and in view of earlier references (§4(iv), §5(ii)) to the possible theoretical interpretations of these characteristics, only a very brief survey is necessary.

Alloys. The permanent magnet alloys which were brought into use before 1930 were all of the quench-hardened type, and usually contained about 1 per cent of carbon. This applies to the carbon, tungsten, chromium and cobalt steels, with $(BH)_{max}$ values of up to about 0·2, 0·3, 0·3 and 1·0 respectively. The alloys have a complex martensitic structure, and it may reasonably be presumed that there are large heterogeneities of internal stress with which the relatively high coercivities (ranging from about 50 for carbon steel to 250 for 35 cobalt steel) compared with those of annealed materials may be associated. There are, however, so many uncertainties and difficulties that no quantitative correlation of the magnetic properties with the structural characteristics has yet been made.

The discovery by Mishima (1931) that carbon-free iron–nickel–aluminium alloys of approximate composition Fe_2NiAl, suitably heat treated, had a coercivity of about 500 opened a new era in the history of permanent-magnet materials.

Additions of cobalt and copper were found to improve the properties, and an alloy introduced in 1934 under the name 'alnico' (Al 10, Ni 18, Co 13, Cu 6), with a $(BH)_{max}$ value of 1·6, remained as the 'best' commercial material until the introduction some years later of a number of anisotropic alloys, produced by cooling in a magnetic field. These alloys contain more cobalt than ordinary alnico. The trade varieties (alcomax, ticonal, alnico, with sub-varieties designated by letters or numbers) all contain Fe, Ni, Al, Co and usually Cu, with various additions, such as titanium and, recently, niobium. Alcomax II, of which demagnetizing curves are shown in Figure 5·2, is reasonably typical, with approximate composition Al 7, Ni 11, Co 25, Cu 6. The $(BH)_{max}$ values range from about 3·5 to 5·5 as averages, which may be exceeded in individual specimens. All these alloys are brittle, and must be cast and ground; alternatively, and with advantages for some purposes, the sintering techniques of powder metallurgy may be used, and are said to give magnetic properties within 10 per cent of those of the equivalent cast alloy. A number of malleable and ductile materials have been produced, among which cunife (60 Cu, 20 Ni, 20 Fe) and vicalloy (52–57 Co, 8–14 V, bal. Fe) may be mentioned, with $(BH)_{max}$ values of about 1·5, and up to 3·5 respectively.

The various alloys are, of course, only in part specified by the approximate composition. The phases present in alloys with optimum magnetic properties have been determined in a few cases by x-ray and other methods, but even here the knowledge is usually incomplete, and definite knowledge of the states of aggregation of the phases is almost entirely lacking. The fuller specification of the alloys is, therefore, necessarily indirect, through statements, in sufficient detail, of the heat treatments (which may include cooling in a magnetic field) and other treatments (e.g. cold reduction) which are required for the development of the specified magnetic properties.

The x-ray and magnetic work by Bradley and Sucksmith respectively on the ternary Fe–Ni–Al alloys has already been summarized (§4(iv) and I, pp. 106–8). The equilibrium state of the alloy of approximate composition Fe_2 Ni Al at high temperatures is single-phase, at room temperature two-phase, the two phases both being body-centred but of slightly different lattice spacings. The phase separation may be represented by

$$3{\cdot}25\ Fe_{50}Ni_{25}Al_{25} = Fe_{95}Ni_{2{\cdot}5}Al_{2{\cdot}5} + 2{\cdot}25\ Fe_{30}Ni_{35}Al_{35};$$

the two phases, it may be noted, are approximately Fe and Fe Ni Al. The specific saturation intensity of the iron-rich phase is about 212, that of the second phase about 61. The single-phase state is retained by quenching, and the two-phase state is produced by long annealing. Neither of these treatments results in an alloy of high coercivity, for which an intermediate rate of cooling is necessary. In the intermediate state, it was suggested (Bradley and Taylor 1938b) that there was an incipient precipitation of the iron-rich phase, with the formation of submicroscopic iron-rich regions. The evidence is consistent with the formation of more strongly ferromagnetic islands, below the critical size for domain boundary formation to be possible, in a less ferromagnetic matrix, and, as discussed by Stoner and Wohlfarth (1948) (cf. §4(iv)), high coercivities can then be accounted for readily in terms of the rotational processes occurring in anisotropic single-domain particles. Whether shape, magneto-crystalline or strain anisotropy is predominant must at present be left open, and it must be remembered that, since the matrix is itself ferromagnetic and some of the islands formed may be

relatively large, processes other than rotations in single domains are involved. None the less, it seems probable that the single-domain processes, which have been fully examined theoretically, play a centrally important part in determining the high-coercivity characteristics of many permanent-magnet alloys of the dispersion hardening type. It should be noted that although the proposed mechanism requires that the individual particles segregated should be small, the aggregate volume occupied by them may be a large fraction of the total volume.

In more recent work Bradley (1949 a, b) has made a comprehensive microscopic study of alloys in the Fe–Ni–Al system. A comprehensive x-ray and metallographical study of this system has also been made by Kiuti (1941), with results in substantial agreement with those of Bradley.

The anisotropic alloys which have been most fully studied are probably those of the alcomax type (see Figure 5.2). The heat treatment required to give optimum properties has been fully investigated by Oliver and Hadfield (1946). For a particular alcomax alloy (11·5 Ni, 7·5 Al, 25·0 Co, 6·0 Cu, bal. Fe) the best properties are obtained by cooling from a 'solution temperature' (in the range 1200° to 1300° c. or 875° to 925° c.) at a controlled rate (about 1·6 deg. sec^{-1}) in a magnetic field of about 4500 oersteds (which must be maintained from above the Curie temperature at about 875° c. to about 650° c.), and subsequently tempering at about 600° c. for 15 hours or more. It may be noted that any considerable deviation from the temperatures stated, say by 50°, results in a very large deterioration of the final properties. (These technical details are given here to illustrate that much more is needed in the barest specification of a magnetic material than a statement of its composition.) A very extensive 'exploratory' x-ray study of these alloys has been made by Oliver and Goldschmidt (1947), the phase structure having been examined after a wide variety of heat treatments. The x-ray work indicates that the development of high coercivity is associated with incipient segregation in a metastable phase as in the simpler ternary alloys, but the totality of structures and transformations revealed as posssible in the system is of bewildering complexity. The significance of many of the findings will hardly become clear until a better understanding is gained of the simpler ternary and binary systems.

The effect of cooling in a magnetic field is that in the direction in which the field has been applied the remanence, coercivity, and 'fullness' are all increased, and in a perpendicular direction decreased. The demagnetizing curves in the favoured direction and at right angles to it (see Figure 5.2) bear a remarkably close resemblance to those which would arise with single domain particles with easy axes approximately parallel to the field (say the curves for 10° and 80° in Figure 4.3, as compared with the mean curve in Figure 4.4). On the basis of the single-domain mechanism, it is suggested that cooling in a magnetic field favours energetically the formation of segregates with an easy direction of magnetization parallel to the field, independently of how the anisotropy arises (Stoner and Wohlfarth 1947, 1948, Néel 1947 d). Whether or not a field is applied, there must, however, be some relation between the orientation of the segregate and that of the crystal in which it is formed. Measurements of magnetostriction of alcomax by Hoselitz and McCaig (1949) are consistent with the view that, after cooling in a magnetic field, the easiest directions of magnetization (i.e. the directions of the domain magnetization vectors in the absence of a field) are along or near the cube-edge directions which are nearest to the anisotropy axis of the specimen. This is

supported by later measurements of McCaig (1949) on samples with 'columnar' crystals, that is polycrystalline samples in which all the crystals have one axis parallel to a given direction. This particular work has no direct bearing on the question of the detailed mechanism giving rise to high coercivity, but it does provide important additional information which must be taken into account in any attempt to explain more fully the totality of magnetic properties of these alloys.

Since the details of the process of phase segregation have such an important bearing on magnetic properties, mention should be made of some x-ray work by Daniel and Lipson (1943, 1944), following earlier observations by Bradley, on an alloy of approximate composition $Cu_4Ni\ Fe_3$, which, when suitably heat-treated, develops permanent-magnet characteristics. Quenched from 1,000° c., this alloy is single-phase, and gives sharp x-ray diffraction lines. On prolonged annealing at 650° c. it segregates into two distinct face-centred phases with different lattice spacings, and the original lines are doubled. With short annealing periods the sharp lines characteristic of the single-phase state are flanked by diffuse but strong 'side-bands'. It is shown that these side-bands, characteristic of the initial stages of segregation during which the coherence of the original lattice is retained, can be accounted for as arising from a periodic variation of the lattice spacing. The spacing increases where the concentration of the copper atoms increases and conversely. The dependence of the wavelength and amplitude of the variation on the time of annealing was studied in detail. Preliminary magnetic measurements by Sucksmith have indicated that as the wavelength increases (up to something over 100 lattice spacings) the remanence at first increases rapidly, and then falls off slowly, and the coercivity increases more gradually to a fairly well defined maximum (Anon. 1946, Sucksmith 1949 a). These effects have usually been attributed to a variation of internal stress associated with the spacing variation, on the general basis of the Becker–Kersten strain theory (§ 4 (ii)), but no quantitative treatment of the problem appears to have been published. Further work is, however, in active progress on the side-band effect, the importance of which is, of course, not confined to ferromagnetic materials.

A number of illustrative examples have been given of the bearing of x-ray investigations, additional to the purely structural studies, on magnetic problems. It is, however, impracticable to deal adequately here with this theme. For a survey of x-ray work on metals and alloys generally, much of which is directly or indirectly relevant to questions considered in this section, reference may be made to the book by Barrett (1943), and for more recent work to a comprehensive review article by Douglas (1949).

The permanent-magnet alloys which have been considered have been almost exclusively ternary or more complex alloys, most of them being alloys (or prototypes of alloys) of technological interest. Relatively high coercivity is also developed in a number of binary (or nominally binary) alloy systems. Among these are alloys of platinum, which are of particular scientific interest (cf. I, p. 92), even though their commercial development as permanent-magnet materials is somewhat improbable. The platinum–iron alloys have been studied by Graf and Kussmann (1935). For an alloy of composition Pt 78 Fe 22, which corresponds to equal atomic percentages, $(BH)_{max}$ values as high as 3 are obtained (B_r 5800, H_c 1570). Microscopic and x-ray studies of this alloy, after various heat treatments, have been made by Lipson, Shoenberg and Stupart (1941), who conclude, in disagreement with Graf and Kussmann, that the structure is tetragonal. In the

high coercivity state the microstructure shows crystal grains crossed by sets of parallel bands, which are interpreted as twinning bands due to the breakdown of the high-temperature face-centred cubic structure. For the equiatomic platinum–cobalt alloy, Pt 77 Co 23, coercivities as high as 4,000 have been obtained (Jellinghaus 1936), and values of $(BH)_{max}$ greater than 4 are quoted (cf. Hoselitz 1944, Oliver and Hadfield 1948, see also Douglas 1949, Sucksmith 1949 a). From what is known about the structure of the platinum alloys in the high-coercivity state, the occurrence of large internal-stress variations seems reasonably certain; but whether these stress variations alone can account for the coercivity through the disperse field effect (§ 4 (ii)) is still an open question.

Investigations have recently been made of binary alloys of a different type in the iron–molybdenum system by Sucksmith and Margerison (1949). Over the range of compositions from Mo 8 to Mo 35 the alloy at high temperatures consists of a solid solution of Mo in Fe, which is retained by quenching. Subsequent heat treatment causes the precipitation of a non-magnetic phase of approximate composition Fe_3Mo_2. The alloys studied contained Mo 10, 15 and 20. Through the development of a method of determining magnetization and hysteresis curves at temperatures up to over 800° C., it was possible not only to examine the magnetic properties of the alloys at room temperature after various heat treatments, but also to follow the manner in which the saturation magnetization, the remanence and the coercivity varied with temperature. From other evidence it is probable that the non-magnetic phase precipitates in the form of plate-like structures, giving layer-like or cell-like non-magnetic boundaries to the thus separated regions of the more abundant ferromagnetic phase. If it is assumed that the fairly high coercivities (up to over 100) are due to stress variations, an approximate theoretical estimate of the magnitudes of these can be made; the values so obtained appear to be considerably higher than is compatible either with the sharpness of x-ray diffraction lines, or with the small effect of prolonged heating at 700° C. on the value of the coercivity. It is therefore suggested that rotational hysteresis in single-domain regions contributes to the observed effects, and also, since over a considerable range there is a linear variation of H_c with I_0, that shape anisotropy is primarily involved. On this basis, a qualitative explanation can be given of some of the main effects observed. It would not be expected that the experimental results would agree, in any completeness, with the theoretical results (§ 4 (iv)) for an assembly of non-interacting single-domain particles. It is, indeed, remarkable that the unelaborated theoretical treatment applies as closely as it appears to do to an alloy in which the aggregate volume of the ferromagnetic 'particles' considerably exceeds that of the matrix, and in which only a proportion of the particles are likely to be single-domain. Further experimental investigations, such as are in progress, on alloys of the iron–molybdenum type, and of complementary alloys in which the precipitation of a ferromagnetic phase from a non-ferromagnetic solid solution can be controlled, offer great promise of leading to a better understanding of the phenomena underlying the development of high coercivity in alloys.

Micro-powder magnets. Reference has already been made (§ 4 (iv)) to the experimental and theoretical work of Néel and his collaborators on magnetic materials made up of very fine particles, which may be presumed to be mainly single-domain, and to the development of these materials for permanent magnets. The materials are essentially different from powder agglomerates such as sintered alcomax, with constituent grains neither isolated nor, necessarily, very small.

The distinctive description of the new materials as 'micro-powder' or 'ultra-fine powder' materials is therefore appropriate. Methods of preparation which have been described in patent specifications include the precipitation of suitable salts (such as formates, acetates and oxalates), or mixed salts (e.g. ferrous iron and cobalt salts) from solution, followed by decomposition of the salts at temperatures lower than 600° c., treatment in various atmospheres (reducing in one process, oxidizing in another), and compression, with or without a binding agent, of the powder so formed. Information about the commercial products is not readily accessible, but Oliver and Hadfield (1948) give for micro-powder iron (with 15 per cent Fe_3O_4) a $(BH)_{max}$ value of 1·0 (B_r 5000, H_c 600), and for iron–cobalt (Fe 37, Co 26, Fe_3O_4 22·5) a value of 1·7 (B_r 7500, H_c 500). The values are comparable with those for ordinary alnico.

Extensive investigations of these micro-powder materials have been made by Néel, Weil and others. No full collected account of this work seems to have been published, but various aspects of it are presented in a number of short notes. In the first of these to which reference has already been made (§ 4 (iv)), Néel (1947 b) considers the critical size for single-domain particles, and examines the properties of assemblies of such particles with magneto-crystalline characteristics similar to those of iron. He shows that if the anisotropy were of magneto-crystalline origin, the mean coercivity, ${}_IH_{cm}$, for random orientation, would be approximately $0{\cdot}64K/I_0$, that is about 160 for iron. Since coercivities of about 1000 have been observed, he concludes that shape anisotropy is also involved. In a second paper (1947 c) shape anisotropy is further considered, and the effect of the mutual magnetic interaction of the particles is discussed. An important relation is given, though not derived in detail, for the mean coercivity of a powder, namely

$$ {}_IH_{cm} = ({}_IH_{cm})_0\{1-(\rho/\rho_0)\}, \qquad \ldots\ldots(5.8) $$

where ρ is the density of the powder, ρ_0 that of the metal in bulk, and $({}_IH_{cm})_0$ a limiting coercivity for effectively non-interacting particles ($\rho \to 0$). With shape anisotropy alone for an assembly of randomly orientated similar prolate spheroids $({}_IH_{cm})_0$ has the value $0{\cdot}48\ (N_b - N_a)I_0$, the maximum value being $0{\cdot}96\pi I_0$ (cf. § 4 (iv)). For a particular, somewhat artificial, distribution of eccentricities of the spheroidal particles, which, however, probably gives a fair representation of an actual powder, Néel obtains the result

$$ ({}_IH_{cm}) = 0{\cdot}26\pi I_0. \qquad \ldots\ldots(5.9) $$

The relation (5.8) has been verified experimentally with fair approximation by Weil (1947) for non-agglomerated powders (held in paraffin) of iron and iron–cobalt with effective values of ρ/ρ_0 ranging from about 0·5 to 0·1. The extrapolated limiting values of ${}_IH_{cm}$ were about 670 and 1000 for the iron and iron–cobalt powders respectively, as compared with $({}_IH_{cm})_0$ values, from (5.9), of about 1400 and 1540. With specially prepared powder materials actual ${}_IH_c$ values of up to and about 1080 and 1230 have been obtained for iron and iron–cobalt respectively.

The actual size of the iron powder particles has been estimated by Weil (1948 a) from the 'heat of wetting' in organic liquids, as compared with that measured for coarser carbonyl-iron powders whose particle sizes are determinable microscopically. The estimated mean diameter of $4{\cdot}5 \times 10^{-6}$ cm. is regarded

as an upper limit (in view of the tendency of the particles to cohere in groups), and is rather larger than the value indicated by x-ray measurements (about $3{\cdot}0 \times 10^{-6}$). These values are satisfactorily consistent with the theoretical estimates of size for single-domain particles (§ 3 (iv)).

Measurements of the variation with temperature of the coercivity of these materials are of value as giving some indication of the origin of the anisotropy. If the anisotropy is primarily of magneto-crystalline origin, ${}_IH_c$ should vary as K/I_0 and, if it is primarily due to shape, as I_0. Work on nickel powders (Weil and Marfoure 1947, Weil 1949) covering in all a temperature range from about $-250°$ c. to the Curie point (about 360° c.) indicates that at very low temperature the magneto-crystalline effect is predominant (i.e. ${}_IH_c$ varies approximately as K/I_0), but that over most of the range shape anisotropy is equally or more important. For cobalt powders, examined by Weil, Marfoure and Bertaud (1948), the effects are complicated by the presence of both hexagonal and cubic forms. The interesting results obtained indicate that, at room temperature, in hexagonal cobalt the magneto-crystalline and shape effects are of comparable importance, and in cubic cobalt the shape effect predominates. Four ferronickel powders, with from 10 to 50 per cent Fe, have been investigated by Weil (1948 b) over a range of effective densities as for the iron powders. For these powders the anisotropy is due almost entirely to shape. The coercivity varies with density in the way indicated by (5.8), and the ratio of the limiting value of ${}_IH_c$ to I_0 was approximately 0·5 in all cases, the limiting values of ${}_IH_c$ ranging from about 320 to 630.

Further investigations on micro-powder materials are in active progress. This brief summary of the published work must suffice to give a general impression of what has been achieved. From a theoretical standpoint it would seem that permanent-magnet characteristics appreciably better than those that have been specified might be obtainable, but the practical difficulties are formidable. Whether or not these materials eventually acquire a value for technological applications comparable with that of the more recent permanent-magnet alloys, they are of outstanding scientific interest in many ways, particularly in relation to the account given here of permanent-magnet materials generally, in manifesting so directly the occurrence of rotational hysteresis effects in single-domain particles.

§ 6. CONCLUSION

In this report an account is given of scientific work over a period of some sixteen years on the behaviour of ferromagnetic materials in low and moderate fields, and in particular of those investigations, both experimental and theoretical, which have contributed to a better knowledge and understanding of just 'what is happening inside a ferromagnetic when it is magnetized'. Ten years ago it would have been easy to gain the impression from much that was written on ferromagnetism that, except for a few refinements, a neat co-ordination, almost a codification, had been effected of all those properties summarized in magnetization and hysteresis curves. Such an impression will hardly have been given by this report. The diversity of the magnetic behaviour of ferromagnetics cannot be unified by a few relatively simple formulae, as has been made abundantly clear by the more recent work which has been reviewed. At the same time, notably

through the work on domain structure and boundary movement processes, the foundations have been laid for a proper analysis and interpretation of the very varied sequences of changes of magnetization accompanying changes of field in different materials, and for a much more satisfactory co-ordination of ferromagnetic characteristics than was previously possible.

It is unnecessary to give here a summary of the report; it is adequately summarized by the table of contents and the outline in the Introduction. The report is long, reflecting the amount of work which has been carried out during the period under review, but neither in fact nor in intention is it exhaustive. Little space has been devoted to work on magneto-mechanical effects, for example, for reasons mentioned earlier (§ 4 (i)), but frequent reference has been made to the results obtained in their bearing on the interpretation of magnetic properties. Some of the later work on magneto-mechanical effects is surveyed by Bates (1950), who also discusses magneto-electric effects; these have not been considered here, as little more would have been possible than an isolated summary of some of the experimental results. The main omission, however, is an account of work on time effects in magnetization, except incidentally in connection with the Barkhausen effect (§ 3 (iii)), and in a few other places. An adequate account would have made the report considerably longer, without adding appreciably to an elucidation of the main problems under consideration Time effects include those diverse phenomena which have been studied under .such names as magnetic viscosity, magnetic after-effect, and magnetic ageing. A discussion of these and other relaxation effects is given by Becker and Döring (1939, pp. 242–67), and there is a tentative analysis of some of them in the book by Snoek (1947); among recent papers in which new contributions are made, and which may also serve as an approach to the earlier work, are those of Snoek (1949) and of Street and Woolley (1949). The effect of macroscopic eddy currents on the time sequence of the change of magnetization after application or removal of a field is usually regarded as in principle calculable, and eddy current effects (including the skin effect) in alternating fields are treated in detail in many standard works. With increasing frequency, microscopic eddy current effects, depending on the details of the localized elementary processes of change of magnetization, become increasingly important; and considerable attention has been devoted recently both to the experimental study, and the theoretical interpretation of the variation with frequency of the reversible permeability, as measured in an alternating field, with or without a superimposed steady field. Although there are a number of unsolved problems in connection with the low-frequency effects, it is the high-frequency behaviour, in particular in the centimetre wavelength range, which is of central interest in relation to the questions discussed in this report. Some account of this work is given by Becker and Döring (1939, pp. 228–42), and there is a comprehensive review by Allanson (1945) of methods of measurement, results obtained, and 'explanatory hypotheses'. Among recent papers may be mentioned those of Birks (1948, 1950), Johnson and Rado (1949) and Millership and Webster (1950).

This report is intended to be reasonably self-explanatory, and in consequence, like the earlier report, it overlaps to some extent books and general articles written during the period under review. A substantial part of the work considered, mainly but not entirely that since about 1939, has, however, not been previously

reviewed in any detail, or had not, at the time of writing. The balance aimed at would have been destroyed if material had been selected merely on the ground that it had not been reviewed before; but in general no more than summary accounts have been given of work of which adequate reviews are readily available. The books on magnetism and on ferromagnetism which are listed in section A of the References provide, it need hardly be said, a much fuller account of the background, and in many cases a fuller account of topics which have only been touched on here. The indebtedness to Becker and Döring's *Ferromagnetismus* (1939) will be apparent from the many references which have been made to it. The debt is not lessened by the fact that more recent work has thrown doubt on the adequacy of the co-ordinating scheme which is so convincingly maintained in that remarkable book, with its wealth of information and illuminating commentary on almost every aspect of the magnetization-curve side of ferromagnetism. In section B of the references, review articles are distinguished by quotation of their titles. Reference has already been made to all of these articles, some of which may be found useful as giving shorter general surveys of the field, and others as giving more detailed reviews of special topics. For collections of data, the review articles on magnetic materials are an indispensable complement to this report (see §5(i)). In the references to original work, an attempt has been made to include all the more substantial papers in the field dealt with, and also the most recent papers of any closely related series. Preliminary communications and notes have been excluded except in special cases. It is believed that practically all the original work will be found at not more than one remove from the papers listed.

In closing this second and concluding report on ferromagnetism, the hope may be expressed that the two reports will enable those who are relatively unfamiliar with the field to gain a clear and reasonably detailed impression of its many-sided interest. The interest is not localized in ferromagnetic materials; for the conclusions drawn as to the internal structure (crystalline, sub-crystalline or atomic) of these materials from a study of the readily investigated ordinary magnetic characteristics may often be relevant to the structure of solids generally. The reports may also be useful to those actively engaged in the investigation of ferromagnetism, and in the development and application of ferromagnetic materials. That the reports are long is partly due to there having been no previous articles concerned exclusively with ferromagnetism in the *Reports on Progress in Physics*. The background provided may make less formidable the task of future writers of reports in this general field in enabling them to deal more readily with particular aspects of the subject, or with recent advances in those directions in which notable progress may have been made.

ACKNOWLEDGMENTS

It is a pleasure to acknowledge the assistance I have had in the preparation of this report from Dr. E. P. Wohlfarth, now at Imperial College, University of London, and from Dr. P. Rhodes and Mr. B. A. Lilley, at the University of Leeds. I also take this opportunity of thanking all those who have given me information in conversation or in correspondence about magnetic work in progress, or who have sent reprints. Acknowledgment for the photographs, from which the plates were prepared, is made in the text.

REFERENCES

References are not included for work published before 1934. For the earlier work mentioned in the report, references may be found in Stoner, *Magnetism and Matter* (1934), or in other books listed in Section A.

A. *Books*

v. Auwers, O., 1938, *Magnetische und elektrische Eigenschaften des Eisens und seiner Legierungen* aus Gmelins Handbuch der anorganischen Chemie (Berlin : Verlag Chemie G.m.b.H.).

Barrett, C. S., 1943, *Structure of Metals* (New York : McGraw-Hill).

Bates, L. F., 1950, *Modern Magnetism* (Cambridge : University Press).

Becker, R. (ed.), 1938, *Probleme der technischen Magnetisierungskurve* (Berlin : Springer).

Becker, R., and Döring, W., 1939, *Ferromagnetismus* (Berlin : Springer). (Photolithoprint reproduction, 1943, Ann Arbor, Michigan : Edwards Brothers.)

Bitter, F., 1937, *Introduction to Ferromagnetism* (New York : McGraw-Hill).

Brailsford, F., 1948, *Magnetic Materials* (Methuen's Monographs on Physical Subjects). (London : Methuen).

Hoselitz, K., 1950, *Ferromagnetic Properties of Metals and Alloys* (Oxford : Clarendon Press).

Kersten, M., 1943, *Grundlagen einer Theorie der ferromagnetischen Hysterese und Koerzitivkraft* (Leipzig : Hirzel). (Photo-lithoprint reproduction, 1946, Ann Arbor, Michigan : Edwards Brothers).

Snoek, J. L., 1947, *New Developments in Ferromagnetic Materials* (Amsterdam : Elsevier).

Stoner, E. C., 1934, *Magnetism and Matter* (London : Methuen).

Stoner, E. C., 1948, *Magnetism* (Methuen's Monographs on Physical Subjects) (London Methuen).

Sykes, C. (ed.), 1938, *Magnetism* (lectures delivered before the Manchester and District Branch of the Institute of Physics) (London : Institute of Physics).

B. *Papers*

The titles are given of papers which are primarily review articles.

Adcock, F., and Bristow, C. A., 1935, *Proc. Roy. Soc.* A, **153,** 172–200.

Akulov, N. S., 1936, *Z. Phys.*, **100,** 197–202.

Allanson, J. T., 1945, ' The permeability of ferromagnetic materials at frequencies between 10^5 and 10^{10} c/s. ', *J. Instn. Elect. Engrs.*, **93** (Pt. III), 247–55.

Anon., 1946, ' Permanent magnet materials.' Summary of recent work carried out at the Cavendish Laboratory, Cambridge, under the direction of Professor Sir Lawrence Bragg, *B.E.A.I.R.A.*, Report N/T 35.

Anon., 1949, ' The magnetic properties of nickel-iron alloys ' (London : Mond Nickel Co.).

Bates, L. F., 1945, *Phil. Mag.*, **36,** 297–318.

Bates, L. F., and Edmondson, A. S., 1947, *Proc. Phys. Soc.*, **59,** 329–43.

Bates, L. F., and Harrison, E. G., 1948 a, *Proc. Phys. Soc.*, **60,** 213–25; 1948 b, *Ibid.*, **60,** 225–36.

Bates, L. F., and Healey, D. R., 1943, *Proc. Phys. Soc.*, **55,** 188–202.

Bates, L. F., and Neale, F. E., 1949, *Physica*, **15,** 220–24; 1950, *Proc. Phys. Soc.* A, **63,** 374–88.

Bates, L. F., and Weston, J. C., 1941, *Proc. Phys. Soc.*, **53,** 5–34.

Becker, R., 1948, ' Ferromagnetismus ', *Fiat Review of German Science*, 1939–46, pp. 27–42.

Birks, J. B., 1948, *Proc. Phys. Soc.*, **60,** 282–92; 1950, *Proc. Phys. Soc.* B, **63,** 65–74.

Boothby, O. L., and Bozorth, R. M., 1947, *J. Appl. Phys.*, **18,** 173–76.

Bozorth, R. M., 1936, *Phys. Rev.*, **50,** 1076–81; 1937, ' Directional ferromagnetic properties of metals ', *J. Appl. Phys.*, **8,** 575–88; 1947, ' Magnetism ' (from Encyclopaedia Britannica, Chicago), *Rev. Mod. Phys.*, **19,** 29–86; 1948, *Z. Phys.*, **124,** 519–27; 1949, ' Ferromagnetic domains ', *Physica*, **15,** 207–19.

Bozorth, R. M., and Williams, H. J., 1941, *Phys. Rev.*, **59,** 827–33.

Bradley, A. J., 1949 a, *Physica*, **15,** 175–78; 1949 b, *J. Iron Steel Inst.*, **163,** 19–30.

BRADLEY, A. J., and TAYLOR, A., 1938 a, *Proc. Roy. Soc.* A, **166,** 353–75; 1938 b, *Magnetism,* 89–102 (*see* Sykes, C., Section A).

BRAILSFORD, F., 1938, *J. Instn. Elect. Engrs.*, **83,** 566–75; 1948, ' Electrical sheet steels ', *Ibid.*, **95,** I, 522–31.

BRAILSFORD, F., OLIVER, D. A., HADFIELD, D., and POLGREEN, G. R., 1948, ' Magnetic materials '. A review of progress, *J. Instn. Elect. Engrs.*, **95,** I, 522–43.

BROOKS, H., 1940, *Phys. Rev.*, **58,** 909–18.

BROWN, W. F., 1937, *Phys. Rev.*, **52,** 325–34; 1938 a, *Ibid.*, **53,** 482–91; 1938 b, *Ibid.*, **54,** 279–87; 1939, *Ibid.*, **55,** 568–78; 1940 a, ' Ferromagnetic domains and the magnetization curve ', *J. Appl. Phys.*, **11,** 160–72; 1940 b, *Phys. Rev.*, **58,** 736–43; 1941, *Ibid.*, **60,** 139–47; 1945 ' Virtues and weaknesses of the domain concept ', *Rev. Mod. Phys.*, **17,** 15–19.

BRUKHATOV, N. L., and KIRENSKI, L. V., 1937, *Phys. Z. Sowjet*, **12,** 602–9.

BUSH, H. D., and TEBBLE, R. S., 1948, *Proc. Phys. Soc.*, **60,** 370–81.

CIOFFI, P. P., 1934, *Phys. Rev.*, **45,** 742.

CIOFFI, P. P., WILLIAMS, H. J., and BOZORTH, R. M., 1937, *Phys. Rev.*, **51,** 1009.

CLASH, R. F., and BECK, F. J., 1935, *Phys. Rev.*, **47,** 158–65.

CLEAVES, H. E., and HIEGEL, J. M., 1942, *J. Res. Nat. Bur. Stand., Wash.*, **28,** 643–67.

DANIEL, V., and LIPSON, H., 1943, *Proc. Roy. Soc.* A, **181,** 368–78; 1944, *Ibid.*, **182,** 378–87.

DE BARR, A. E., 1949, *J. Instn. Elect. Engrs.* **96,** II, 719–28.

DILLINGER, J. F., and BOZORTH, R. M., 1935, *Physics*, **6,** 279–91.

DÖRING, W., 1936, *Z. Phys.*, **103,** 560–82; 1938, *Ibid.*, **108,** 137–52.

DÖRING, W., and HAAKE, H., 1938, *Phys. Z.*, **39,** 865–68.

DOUGLAS, A. M. B., 1949, ' X-ray work in some fields of research on metals and alloys, 1939 to 1946 ', *J. Iron Steel Inst.*, **162,** 300–15.

DRIGO, A., and PIZZO, M., 1948, *Nuovo Cim.*, **5,** 196–206.

ELMORE, W. C., 1937, *Phys. Rev.*, **51,** 982–88; 1938 a, *Ibid.*, **53,** 757–64; 1938 b, *Ibid.*, **54,** 309–10; 1938 c, *Ibid.*, **54,** 1092–95; 1939, *J. Appl. Phys.*, **10,** 724–27; 1940 a, *Ibid.*, **11,** 797–99; 1940 b, *Phys. Rev.*, **58,** 640–42; 1942, *Ibid.*, **62,** 486–93.

v. ENGEL, A., and WILLS, M.S., 1947, *Proc. Roy. Soc.* A, **188,** 464–84.

GERLACH, W., 1938, *Probleme der technischen Magnetisierungskurve*, pp. 141–56 (*see* Becker, R. Section A).

GRABBE, E. M., 1940, *Phys. Rev.*, **57,** 728–34.

GRAF, L., and KUSSMANN, A., 1935, *Phys. Z.*, **36,** 544–51.

GUGGENHEIM, E. A., 1936, *Proc. Roy. Soc.* A, **155,** 49–101.

GUGGENHEIMER, K. M., HEITLER, H., and HOSELITZ, K., 1948, *J. Iron Steel Inst.*, **158,** 192—99.

GUILLAUD, C., 1949, *C.R. Acad. Sci., Paris*, **229,** 818–19.

HAAKE, H., 1939, *Z. Phys.*, **113,** 218–33.

HARDY, T. C., and QUIMBY, L., 1938, *Phys. Rev.*, **54,** 217–33.

HAWORTH, F. E., 1938, *Phys, Rev.*, **54,** 693–98.

HEAPS, C. W., 1936, *Phys. Rev.*, **50,** 176–79; 1941, *Ibid.*, **59,** 585–87.

HOBSON, P. T., and OSMOND, W. P., 1948, *Nature, Lond.*, **161,** 562–63.

HOLSTEIN, T., and PRIMAKOFF, H., 1941, *Phys. Rev.*, **59,** 388–94.

HONDA, K., MASUMOTO, H., and SHIRAKAWA, Y., 1935, *Sci. Rep. Tôhoku Univ.*, **24,** 291–310.

HOSELITZ, K., 1944, *Phil. Mag.*, **35,** 91–102; 1946, ' Modern hard magnetic materials ', *J. Sci. Instrum.*, **23,** 65–71.

HOSELITZ, K., and MCCAIG, M., 1949, *Proc, Phys. Soc.* B, **62,** 163–70.

JAANUS, R., and SCHUR, J., 1937, *Phys. Z. Sowjet*, **12,** 383–88.

JELLINGHAUS, W., 1936, *Z. techn. Phys.*, **17,** 33–36.

JOHNSON, M. H., and RADO, G. T., 1949, *Phys. Rev.*, **75,** 841–64.

KAHAN, T., 1938, *Ann. Phys., Paris*, **9,** 105–76.

KÄLLBÄCK, O., 1948, *Ark. Mat. Astr. Fys.*, **34 B,** No. 17, 1–6.

KAYA, S., 1934 a, *Z. Phys.*, **89,** 796–805; 1934 b, *Ibid.*, **90,** 551–58; 1938, *J. Fac. Sci. Hokkaido Univ.*, **2,** 29–53.

KAYA, S., and SEKIYA, J., 1935, *Z. Phys.*, **96,** 53–61.

KERSTEN, M., 1938 a, *Probleme der technischen Magnetisierungskurve*, pp. 42–72 (*see* Becker, R., Section A); 1938 b, *Phys. Z.*, **39,** 860–65; 1943, *Ibid.*, **44,** 63–77; 1948, *Z. Phys.*, **124,** 714–41.

KIRKHAM, D., 1937, *Phys. Rev.*, **52,** 1162–67.
KITTEL, C., 1946, *Phys. Rev.*, **70,** 965–71; 1948, *Ibid.*, **73,** 810–11; 1949, 'Physical Theory of Ferromagnetic Domains', *Rev. Mod. Phys.*, **21,** 541–83.
KIUTI, S., 1941, *Sci. Rep. Tôhoku Univ.*, **29,** 742–94.
KLEIS, J. D., 1936, *Phys. Rev.*, **50,** 1178–81.
KONDORSKI, E., 1937, *Phys. Z. Sowjet*, 11, 597–620; 1938, *Phys. Rev.*, **53,** 1022.
KÖNIG, H., 1948, *Optik*, **3,** 101–19.
KÖSTER, W., 1948, *Z. Phys.*, **124,** 545–65.
KUSSMANN, A., 1935, 'Stand der Forschung und Entwicklung auf dem Gebiet der ferromagnetischen Werkstoffe', *Arch. f. Elek.*, **29,** 297–332; 1939, 'Wege und Ergebnisse der ferromagnetischen Werkstoffforschung', *Z.V.D.I.*, **83,** 445–56.
LACOMBE, P., and BERGHEZAN, A., 1949, *Physica*, **15,** 161–69.
LANDAU, L., and LIFSHITZ, E., 1935, *Phys. Z. Sowjet*, **8,** 153–69.
LAWTON, H., 1949, *Proc. Camb. Phil. Soc.*, **45,** 145–56.
LAWTON, H., SHOENBERG, D., and STEWART, K. H., 1947, *B.E.A.I.R.A.*, Report N/T 43.
LAWTON, H., and STEWART, K. H., 1948, *Proc. Roy. Soc.* A., **193,** 72–88.
LEECH, P., and SYKES, C., 1939, *Phil. Mag.*, **27,** 742–53.
LIFSHITZ, E., 1944, *J. Phys. U.S.S.R.*, **8,** 337–46.
LIPSON, H., SHOENBERG, D., and STUPART, G. V., 1941, *J. Inst. Met.*, **67,** 333–40.
LIVENS, G. H., 1945, *Phil. Mag.*, **36,** 1–20; 1947, *Ibid.*, **38,** 453–79; 1948, *Proc. Camb. Phil. Soc.*, **44,** 534–45.
MCCAIG, M., 1949, *Proc. Phys. Soc.* **B 62**, 652–56.
MCKEEHAN, L. W., 1934 a, 'Ferromagnetism in metallic crystals', *Amer. Inst. Min. Met. Eng.*, Tech. Publ. No. 554, pp. 42; 1934 b, *Rev. Sci. Instrum.*, **5,** 265–68; 1937 a, *Phys. Rev.*, **51,** 136–39; 1937 b, *Ibid.*, **52,** 18–30, 527; 1938 a, *Ibid.*, **53,** 301–7; 1938 b, *Ibid.*, **53,** 307–9.
MCKEEHAN, L. W., and ELMORE, W. C., 1934 a, *Phys. Rev.*, **46,** 226–28; 1934 b, *Ibid.*, **46,** 529–31; 1936, *Trans. Amer. Inst. Min. Met. Eng.*, **120,** 236–50.
MCKEEHAN, L. W., PIETY, R. G., and KLEIS, J. D., 1936, *Rev. Sci. Instrum.*, **7,** 494–97.
MASUMOTO, H., 1936, *Sci. Rep. Tôhoku Univ.*, **25** (Honda Vol.), 388–402.
MILLERSHIP, R., and WEBSTER, F. V., 1950, in course of publication.
MONTALENTI, G., 1948, *Nuovo Cim.*, **5,** 154–61.
MURAKAWA, K., 1936, *Proc. Phys.-Math. Soc., Japan*, **18,** 380–401; 1937, *Ibid.*, **19,** 715–33.
NÉEL, L., 1944 a, *J. Phys. Radium*, **5,** 241–51; 1944 b, *Ibid.*, **5,** 265–76; 1944 c, *Cahiers de Phys.*, **25,** 1–20; 1944 d, *Ibid.*, **25,** 21–44; 1946 a, *C.R. Acad. Sci., Paris*, **223,** 141–2; 1946 b, *Ibid.*, **223,** 198–200; 1947 a, *Ann. Univ. Grenoble*, **22,** 299–343; 1947 b, *C.R. Acad. Sci., Paris*, **224,** 1488–90; 1947 c, *Ibid.*, **224,** 1550–51; 1947 d, *Ibid.*, **225,** 109–11; 1948 a, *J. Phys. Radium*, **9,** 184–92; 1948 b, *Ibid.*, **9,** 193–99; 1949, *Physica*, **15,** 225–34.
OKAMURA, T., 1936, *Sci. Rep. Tôhoku Univ.*, **24,** 745–807.
OLIVER, D. A., and GOLDSCHMIDT, H. J., 1947, *B.E.A.I.R.A.*, Report N/T 41.
OLIVER, D. A., and HADFIELD, D., 1946, *B.E.A.I.R.A.*, Reports N/T 36, N/T 40; 1948, 'Permanent-magnet materials', *J. Instn. Elect. Engrs.*, **95,** I, 531–39.
VAN PEPYE, W. F., 1938, *Physica*, **5,** 465–82.
PFEIL, L. B., POLGREEN, G. R., BUCKLEY, S. E., OLIVER, D. A., and GARVIN, S. J., 1947, Symposium on Powder Metallurgy, Section C, 'Magnetic powders and products', *Iron and Steel Inst.*, Special Report, No. 38, pp. 47–72.
PIETY, R. G., 1936, *Phys. Rev.*, **50,** 1173–77.
POLGREEN, G. R., 1948, 'Magnetic powder cores', *J. Instn. Elect. Engrs.*, **95,** I, 540–43.
POLLEY, H., 1939, *Ann. Phys., Lpz.*, **36,** 625–50.
POTTER, H. H., 1934, *Proc. Roy. Soc.* A, **146,** 362–87.
RANDALL, W. F., 1937, 'Nickel-iron alloys of high permeability, with special reference to mumetal', *J. Instn. Elect. Engrs.*, **80,** 647–58.
REINHART, R. A., 1934 a, *Phys. Rev.*, **45,** 420–24; 1934 b, *Ibid.*, **46,** 483–86.
RHODES, P., 1948, *Proc. Leeds Phil. Soc.*, **5,** 116–27.
RICHER, G. C., 1938, 'Electrical sheet steel', *Magnetism*, pp. 19–30 (*see* Sykes, C., Section A) 1944, *J. Iron Steel Inst.*, **150,** 111–60; 1947, *Nature, Lond.*, **160,** 256.
SCHLECHTWEG, H., 1936, *Ann. Phys., Lpz.*, **27,** 573–96.
SCHOLEFIELD, H. H., 1949, *J. Sci. Instrum.*, **26,** 207–9.

SELISSKI, I. P., and ZAIMOVSKI, A. S., 1941, *J. Phys. USSR*, **4**, 563–68.
SHIH, J. W., 1934, *Phys. Rev.*, **46**, 139–42; 1936, *Ibid.*, **50**, 376–79.
SHOENBERG, D., and WILSON, A. J. C., 1946, *Nature, Lond.*, **157**, 548.
SIXTUS, K. J., 1935, *Phys. Rev.*, **48**, 425–30; 1937 a, *Ibid.*, **51**, 870–7; 1937 b, *Ibid.*, **52**, 347–52.
SNOEK, J. L., 1935, *Physica*, **2**, 403–12; 1936, *Ibid.*, **3**, 118–24; 1941, *Ibid.*, **8**, 344–46; 1949, *Ibid.*, **15**, 244–52.
SNOEK, J. L., and LOUWERSE, M. W., 1937, *Physica*, **4**, 257–66.
STEWART, K. H., 1949, *Proc. Camb. Phil. Soc.*, **45**, 296–304.
STONER, E. C., 1935, *Phil. Mag.*, **19**, 565–88; 1936, *Phil. Trans. Roy. Soc.* A, **236**, 165–93; 1937, *Phil. Mag.*, **23**, 833–57; 1938, 'Magnetization curves of ferromagnetics', *Magnetism*, 43–67 (*see* Sykes, C., Section A); 1944, 'Magnetism in theory and practice', *J. Instn. Elect. Engrs.*, **91**, I, 340–49; 1945, *Phil. Mag.*, **36**, 803–21; 1948, 'Ferromagnetism', *Rep. Prog. Phys.*, **11**, 43–112 (London: The Physical Society).
STONER, E. C., and RHODES, P., 1949, *Phil. Mag.*, **40**, 481–522.
STONER, E. C., and WOHLFARTH, E. P., 1947, *Nature, Lond.*, **160**, 650–51; 1948, *Phil. Trans. Roy. Soc.* A, **240**, 599–642.
STREET, R., and WOOLLEY, J. C., 1949, *Proc. Phys. Soc.* **62** A, 562–72.
SUCKSMITH, W., 1939, *Proc. Roy. Soc.* A, **171**, 525–40; 1949 a, 'The principal work on ferromagnetism since about 1938', *J. Iron Steel Inst.*, **163**, 51–60; 1949 b, *B.E.A.I.R.A.*, Report N/T 52.
SUCKSMITH, W., and MARGERISON, T. A., 1949, *B.E.A.I.R.A.*, Report N/T 50.
TAKAGI, M., 1937, *Sci. Rep. Tôhoku Univ.*, **26**, 55–64; 1939, *Ibid.*, **28**, 20–127.
TARASOV, L. P., 1939, *Phys. Rev.*, **56**, 1224–40.
TARASOV, L. P., and BITTER, F., 1937, *Phys. Rev.*, **52**, 353–60.
TEBBLE, R. S., and CORNER, W. D., 1950, in course of publication.
TEBBLE, R. S., SKIDMORE, I. C., and CORNER, W. D., 1950, *Proc. Phys. Soc.* A, **63**, in the press.
TITOW, E. F., 1936, *Phys. Z. Sowjet*, **10**, 337–47.
TOWNSEND, A., 1935, *Phys. Rev.*, **47**, 306–310.
VAN VLECK, J. H., 1937, *Phys. Rev.*, **52**, 1178–98; 1947, 'Quelques aspects de la théorie du magnétisme', *Ann. de l'Inst. Henri Poincaré*, **10**, 57–190.
VONSOVSKI, S. V., 1940, *J. Phys. USSR*, **3**, 83–93.
WEBB, C. E., 1938, 'Recent developments in magnetic materials', *J. Instn. Elect. Engrs.*, **82**, 303–24.
WEIL, L., 1947, *C.R. Acad. Sci., Paris*, **225**, 229–30; 1948 a, *Ibid.*, **227**, 48–50; 1948 b, *Ibid.*, **227**, 1347–49; 1949, *Ibid.*, **228**, 1561–82.
WEIL, L., and MARFOURE, S., 1947, *J. Phys. Radium*, **8**, 358–61.
WEIL, L., MARFOURE, S., and BERTAUD, F., 1948, *J. Phys. Radium*, **9**, 203–7.
WILLIAMS, H. J., 1937 a, *Rev. Sci. Instrum.*, **8**, 55–60; 1937 b, *Phys. Rev.*, **52**, 747–51; 1937 c, *Ibid.*, **52**, 1004–5; 1947, *Ibid.*, **71**, 646–47.
WILLIAMS, H. J., and BOZORTH, R. M., 1939, *Phys. Rev.*, **56**, 837.
WILLIAMS, H. J., BOZORTH, R. M., and SHOCKLEY, W., 1949, *Phys. Rev.*, **75**, 155–78.
WILLIAMS, H. J., and SHOCKLEY, W., 1949, *Phys. Rev.*, **75**, 179–83.
WILSON, A. J. C., 1946, *Proc. Phys. Soc.*, **58**, 21–29.
WOHLFARTH, E. P., 1949, *Proc. Leeds Phil. Soc.*, **5**, 213–23.
YENSEN, T. D., 1937, 'Magnetic materials and their preparation', *Introduction to Ferromagnetism*, 67–125 (*see* Bitter, F., Section A); 1939, 'Magnetically soft materials', *Trans. Amer. Soc. Metals*, **27**, 797–820; 1948, 'Magnetically soft materials', *Metals Handbook*, 587–95 (Amer. Soc. Met.).

RADIO ASTRONOMY

By M. RYLE
Cavendish Laboratory, Cambridge

CONTENTS

PAGE

§ 1. Introduction 185
§ 2. The first astronomical observations at radio wavelengths 187
§ 3. The expression of random radio-frequency fields in terms of equivalent temperatures 188
§ 4. The measurement of small random radio fields 190
The measurement of small radio-frequency noise powers 190
§ 5. The observation of a small source against the background radiation of the galaxy 193
(*a*) The ' pencil ' beam 194
(*b*) ' Interference ' method 195
(*c*) The detection of discrete sources in the galaxy 197
(*d*) The measurement of the diameter and position of a source 199
§ 6. The measurement of polarization 201
§ 7. Experimental observations of solar radiation 203
(*a*) Observations on wavelengths shorter than 1 m 204
(*b*) Observations on wavelengths greater than 1 m 208
§ 8. Experimental observations of galactic radiation 214
(*a*) The general radiation from the galaxy 214
(*b*) The discovery of discrete sources in the galaxy 215
(*c*) The distribution of the discrete sources 218
(*d*) The polarization of the radiation from the discrete sources 219
(*e*) The cause of the fluctuations of intensity 220
§ 9. Theories of solar radiation 222
(*a*) Application of the magneto-ionic theory to the solar atmosphere 223
(*b*) The types of mechanism suggested to account for the emission of radio waves from the sun 225
(*c*) Radiation from the undisturbed sun 226
(*d*) The production of high-energy electrons in the solar atmosphere 228
(*e*) The enhanced radiation from sunspots 229
(i) Theories based on ' equilibrium ' mechanisms 229
(ii) Theories based on the coherent oscillation of a number of electrons 232
(*f*) Detailed examination of the mean electron energy maintained by an electric field 233
(*g*) Theories of the rapid fluctuations associated with sunspots and the sudden large increases associated with chromospheric flares 235
§ 10. Theories of galactic radiation 236
(*a*) Origin in the interstellar gas 236
(*b*) The nature of the sources 237
(*c*) Mechanisms for the emission of intense radiation from the sources 238
(*d*) The origin of the total galactic radiation 239
§ 11. Future developments in radio astronomy 242
(*a*) The study of the solar atmosphere 242
(*b*) The study of stellar envelopes 243
(*c*) The origin of cosmic rays 244
(*d*) The structure of the galaxy 244
Acknowledgments 245
References 245

§1. INTRODUCTION

OUR knowledge of the Universe has been gained almost entirely by the observation and measurement of the electromagnetic radiation which falls on the earth. Until recent years these observations have been restricted to wavelengths in the visual and near-visual regions of the electromagnetic spectrum; although measuring techniques are available for some 40 octaves of the spectrum, astronomical observations have been restricted to about 4 octaves.

The measurement of radiation in this narrow band of wavelengths has made possible deductions concerning the nature and the physical conditions of remote bodies; it is, however, only possible to examine bodies which are capable of emitting, refracting or absorbing the received radiation. Astronomy at visual wavelengths has therefore mainly been concerned with the study of the more highly condensed regions of the Universe. In particular the direct observation of stellar envelopes has usually been prevented by the relatively small intensity of light radiation emitted by them in comparison with that emitted by the more dense but probably cooler surface regions of the star. The direct observations of the solar envelope, which have been made possible by lunar or instrumental

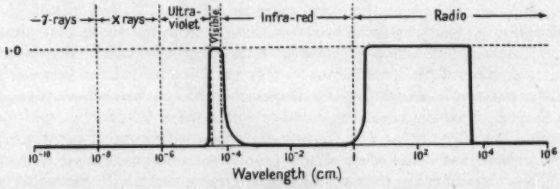


Figure 1. The electromagnetic spectrum, showing the approximate transmission coefficient of the earth's atmosphere.

eclipses of the photosphere, have indicated extremely high temperatures, and it seems probable that information of great importance to astrophysics would be obtained by direct observation of stellar envelopes.

The use of other wavelengths in the electromagnetic spectrum suggests a means of increasing the experimental knowledge of the Universe. Observations on the earth are clearly restricted to those wavelengths which are not appreciably affected by the earth's atmosphere. Unfortunately, over a large proportion of the electromagnetic spectrum for which measuring techniques are available, the atmosphere is almost completely opaque; the extension of the visual and near-visual band to shorter wavelengths is limited by atomic absorption bands, whilst on the long-wave side molecular bands present an almost continuous absorption up to a wavelength of about 1 cm. (Figure 1).

At wavelengths greater than 1 cm., however, the absorption becomes negligible, and it would be possible to make use of wavelengths up to 10–20 metres; beyond this limit the incoming radiation would be reflected back by the ionosphere.

The development of high-altitude rockets has made it possible to carry recording apparatus through a considerable proportion of the earth's atmosphere; in this way observations can be made on wavelengths which would be too heavily absorbed by the whole atmosphere, and an extension of the usable bands of the

spectrum becomes possible. Important observations have already been made in which the solar spectrum was extended on the short-wave side to 2100 A. (Durand, Oberly and Tousey 1949). Whilst some important results are likely to be obtained by the development of further rocket-borne apparatus, it seems unlikely that such brief and expensive experiments will play a major part in the growth of observational astronomy.

Astronomical observations at the surface of the earth are thus restricted to a band about 4 octaves wide in the visual region of the spectrum, and a band about 10 octaves wide between wavelengths of about 1 cm. and about 20 m. Within the last few years it has been found possible to detect radiation in the 'radio' band both from the sun and from galactic bodies.

Before describing the experimental results, it is valuable to consider the differences to be expected between observations at radio wavelengths and light wavelengths. From the astrophysical viewpoint the significance of the radio waves lies in their relatively great absorption, refraction and emission by a low-density ionized gas. The propagation of radio waves and their absorption and emission by a low-density ionized gas will be discussed in § 9 (*a*); it will there be shown that for metre-wave radiation the outer corona of the sun approximates to a black body. A region which emits extremely weak radiation in the visible range of wavelengths can therefore approximate to a 'full' radiator at radio wavelengths. It is therefore clear that the use of radio waves may allow the observation of regions which are too diffuse for observation at visual wavelengths. The examination of the solar corona in this way has already been demonstrated; the direct observation of stellar envelopes might allow important advances in astrophysics. Existing observations have suggested the importance of a more complete knowledge of the mechanisms involved in the solar envelope; the support of prominences, the emission of corpuscular streams and the emission of intense ultra-violet radiation suggest mechanisms of great energy. The possibility that similar mechanisms on a larger scale in other stellar bodies might be responsible for the acceleration of cosmic-ray particles, the fluctuations of the light intensity from certain stars, and other unexplained astrophysical phenomena, indicates the importance of direct observations of stellar envelopes.

On the other hand, the use of radio waves makes it impossible to achieve a resolving power comparable with that available at visual wavelengths and, as will be seen in later sections, many of the limitations of Radio Astronomy are related to the difficulties of separating adjacent sources and of detecting small sources against the general background radiation from the galaxy.

Two excellent survey papers on Radio Astronomy (Hey 1949, Pawsey 1950) have recently been published; the present paper is intended to be complementary to these, and a rather different emphasis has therefore been adopted. Details of the experimental observations have been covered somewhat briefly, whilst the fundamental limitations of the various experimental techniques, and some of the theoretical implications of the present observations, are discussed at some length.

In a survey paper of this nature it is perhaps inevitable that particular emphasis should be placed on the experimental and theoretical work at the Cavendish Laboratory, with which the author has been associated. This is perhaps especially true in the case of the theoretical arguments which are given n §§ 9 and 10, where an attempt has been made to put forward a connected

treatment covering many of the experimental observations; it has not been possible to give such detailed analyses of the theories proposed by other authors, both because little information was available apart from their published papers and because there have been few opportunities of discussing the various difficulties mentioned in §§ 9 and 10.

§ 2. THE FIRST ASTRONOMICAL OBSERVATIONS AT RADIO WAVELENGTHS

The first definite evidence for the existence of measurable radio waves of extra-terrestrial origin was obtained in 1932 by Jansky (1932, 1933) who observed, on a wavelength of 15 metres, random 'noise' signals whose intensity varied with the time of day. At first he attributed these variations to diurnal changes in the ionosphere, but after a few months it was clear that their period was the sidereal day; the origin of the radiation was therefore not connected with the ionosphere or the sun, but must lie outside the solar system. Continued observations showed that the intensity was a maximum when his aerials were directed at the centre of the galaxy.

Later observations by Fränz (1942) on a wavelength of 10 metres confirmed the conclusion that the radiation originated mainly in the plane of the galaxy, and calculations showed that at this wavelength certain regions of the galaxy emitted radiation as if they were at a temperature of about 100,000° K.

Jansky also attempted, without success, to detect similar radiation from the sun; the absence of detectable solar radiation suggested that the galactic radiation was not due to the total emission from the stars, but was possibly caused by the emission from the interstellar gas.

More recently Southworth (1945), using an aerial of considerable resolving power, succeeded in detecting radiation from the sun on much shorter wavelengths between 1 and 10 cm. (where the galactic radiation is far less intense), whilst Hey (1946), working on a wavelength of 4 metres, detected solar radiation whose intensity for short periods during the passage of large sunspots was of greater intensity than the galactic radiation. He was, however, unable to detect the radiation from the undisturbed sun, in the absence of sunspots, owing to the masking effect of the background radiation from the galaxy.

From observations at the shorter wavelengths, Southworth concluded that the intensity corresponded to the emission from a source at a temperature of the same order as that predicted for the photosphere by visual observations (6000° K.). The great intensities observed at a wavelength of 4 metres during the passage of large sunspots could only be accounted for by assuming that the sun radiated as if it had a temperature of about 10^9 deg. K. (Appleton 1945, Appleton and Hey 1946 b).

The early observations of both the galaxy and the sun had shown that if the signal from the aerial was amplified and fed into a loudspeaker or displayed on a cathode-ray tube, it exhibited many of the characteristics of the random 'noise' which is generated in a heated resistor. It was therefore natural to associate the radiation with the emission from a source at a high temperature.

Before describing more recent observations, some fundamental problems in the interpretation and measurement of the small 'noise' powers from the sun and galaxy will be discussed.

§3. THE EXPRESSION OF RANDOM RADIO-FREQUENCY FIELDS IN TERMS OF EQUIVALENT TEMPERATURES

Observations in Radio Astronomy basically involve the measurement of the random radio-frequency power which can be abstracted from an aerial system of known directivity. The methods by which very small radio-frequency powers may be measured will be discussed in §4, and consideration will now be given to the interpretation of the results in terms of the physical conditions of emission in the source.

The radiation falling on the aerial system may originate in a diffuse source, occupying a solid angle large compared with that over which the aerial system is sensitive, or it may be emitted by a source subtending a small solid angle whose magnitude may or may not be known. In practice radiation from both types of source may be incident simultaneously, and methods for distinguishing the components will be necessary. These methods will be described in §5, and for the present it will be assumed that only one type of source is present at a time.

Since the observed radiation has the characteristics of random noise, the power received will be proportional to the bandwidth of the measuring apparatus, and in describing the power emitted by a source it would be necessary to specify the bandwidth over which the power was observed. In measurements of random noise in circuits, this difficulty has been overcome by the use of an 'equivalent temperature' to describe the noise power, a method which has the additional advantage of avoiding complications due to impedance mismatch. The concept of 'equivalent temperature' provides a convenient method of describing the power from an aerial system in a random noise field.

According to Nyquist (1928), the available power which may be abstracted from a resistance at a temperature T is given by the expression $\boldsymbol{k}T\delta f$, where $\boldsymbol{k}$ is Boltzmann's constant and δf is the frequency band over which the power is measured. Burgess (1941, 1946) and Lehmann (1946) have extended this idea to the radiation resistance of an aerial, and have shown that the power available from an aerial in a black-body enclosure at a temperature T is also given by $\boldsymbol{k}T\delta f$. If such an aerial were connected by a transmission line to a resistance, equilibrium would be established when the temperature of the resistance was equal to that of the enclosure; by adjusting the temperature of the resistance until there was no net flow of power, it would thus be possible, theoretically, to measure the temperature of the enclosure.

Simple thermodynamic arguments indicate that the same conditions will hold if the aerial system is directive; equilibrium will still be attained when the temperatures of the resistance and enclosure are equal. Let the receptivity of the aerial in a direction (θ, ϕ) be given by $f(\theta, \phi)$; then the effective solid angle for reception may be written $\Omega = \iint f(\theta, \phi) \cos\phi \, d\theta \, d\phi$, and the power gain of the aerial relative to an isotropic radiator is $4\pi/\Omega$. If now the aerial is not mounted in an enclosure, but is directed at a black body which subtends a solid angle greater than the solid angle for reception, Ω, equilibrium is still established when the resistance is at the same temperature as the black body; if, however, the black body subtends a solid angle smaller than Ω, the power from the aerial will be reduced proportionally.

It is now possible to apply these arguments to the reception of random radio waves from galactic sources. Suppose first that radiation is received from a

diffuse source, which subtends a solid angle greater than Ω. If the noise power measured by the receiver corresponds to that produced by a matched resistance at a temperature T_A, it would be possible to connect the aerial to a resistance at this temperature and thus produce an equilibrium condition. The 'equivalent aerial temperature', T_A, now gives a measure of the emission from the source, since if the source were a perfect black-body radiator over the frequency band of the receiver, equilibrium would be established when T_A was equal to the temperature of the source.

A comparison of the power available from the aerial with that from a heated resistance of the same impedance thus makes it possible to determine an 'equivalent source temperature' for a diffuse source which occupies a larger solid angle than that of the aerial (Ω).

If now radiation is received from a source which subtends a solid angle ω that is smaller than Ω, equilibrium between the aerial and a heated resistance will be established when the available powers from the resistance and aerial are equal. The available power from the aerial is, however, reduced in the ratio ω/Ω relative to that obtained with a large source; equilibrium between aerial and resistance will therefore now be established when the emission from the source is greater (by a factor Ω/ω) than that from a source at a temperature T_A. For black-body radiation at radio wavelengths the Rayleigh–Jeans law is applicable and the emission per unit frequency band is proportional to the temperature; the emission from the source therefore corresponds to that from a black-body radiator at a temperature T_S, where $T_S = (\Omega/\omega)T_A$.

By comparing the available power from the aerial with that from a heated resistance, and knowing the solid angle for reception, it is therefore possible to deduce an 'effective source temperature' T_S for a source subtending a smaller solid angle ω.

If radiation is received from a small source which subtends an unknown solid angle (such as the 'point' sources in the galaxy) it is not possible to deduce an effective source temperature, and the intensity of the emission can only be described in terms of the flux produced, within a given frequency band, at the earth. The available power produced in the aerial by a given incident flux is determined by the 'effective area' of the aerial system, a quantity which is related to the solid angle for reception by the expression $A = \lambda^2/\Omega$, where λ is the wavelength (see, for example, Smith 1949).

It is important to note that when the emission from a source is described in terms of the power flux at the earth, it is also necessary to specify the polarization of the aerial system used. Thus if a linearly- or circularly-polarized aerial is used to measure randomly-polarized radiation, or if a linearly-polarized aerial is used to measure circularly-polarized radiation, the observed power flux will only be half the total incident flux. This complication does not arise in the case where the emission is described in terms of an equivalent temperature.

It would, of course, also be possible to describe the emission from diffuse sources and from sources which subtend a small but known solid angle in terms of the flux produced at the earth in a given frequency band from unit solid angle of the source. Apart from the inconvenience of such a unit, the method of measuring the power from the aerial usually involves comparison with some local source of noise power, and in this way the idea of an equivalent aerial temperature is introduced. The concept of an equivalent source temperature

also has advantages in considering the processes of emission from a body, although it should be emphasized here that it is used merely as a convenient unit of the emission; it is not implied that galactic sources of radio waves necessarily radiate as if they were black bodies, as will be obvious in later sections where theories of the emission are discussed.

§4. THE MEASUREMENT OF SMALL RANDOM RADIO FIELDS

Early observations at wavelengths shorter than 10 cm. had shown that the emission from the sun was of the same order of magnitude as that expected from a source at a temperature of 6,000° K., whilst no detectable radiation was observed from the galaxy. Observations at wavelengths of about 4 metres had shown that during periods of marked sunspot activity the intensity of the radiation corresponded to the emission from a source subtending the same solid angle as the sun, and at a temperature of about 10^9 deg. K. Radiation of considerable intensity from the galaxy was also observed, and at a longer wavelength of 10 metres corresponded to the emission from a source at a temperature of about 10^5 deg. K. No radiation had been detected at these longer wavelengths from the undisturbed sun, partly owing to insufficient sensitivity of the receiving equipment, and partly due to the masking effect of the galactic radiation when using aerials of low resolving power (having a large Ω).

It is clearly important to be able to observe the sun at these longer wavelengths at all times, and not only at times of abnormal activity. By the development of more sensitive receiving apparatus, and the construction of aerials of considerable resolving power, it has become possible not only to observe the radiation from the undisturbed sun, but also to detect discrete sources of radio waves in the galaxy.

The problem of measuring the small random radio-frequency powers which can be abstracted from the aerial, and the design of aerials of large resolving power to enable the radiation from a small source to be distinguished from that due to the general galactic background, will now be discussed.

The measurement of small radio-frequency noise powers.

In order to measure the small powers which are available from the aerial system (and which will be described in terms of the equivalent aerial temperature T_A), it is necessary to use an amplifier whose input circuit itself gives rise to a random noise power which may be considerably greater than that from the aerial. The noise power due to Johnson and shot effects which are inherent in the input circuit of the receiver may be described in terms of an equivalent temperature T_R. In the absence of an input from the aerial, the output indication of the amplifier will fluctuate in a random manner about a mean value which corresponds to the input power associated with T_R. The instantaneous value of the output indication may have any magnitude, but the mean value obtained from a large number, n, of instantaneous observations will depart from the true mean value by an amount proportional to $1/\sqrt{n}$. If the bandwidth of the circuits preceding the output indicator is B, it is possible to obtain B independent observations per second. If the time-constant of the output indicator itself is now increased, so that it corresponds to a bandwidth b, it will record the average output over a period of $1/b$ seconds, so that it records

effectively the average of B/b independent observations. It can therefore be seen that the reading of the output indicator will fluctuate about the true mean value by an amount proportional to $\sqrt{(b/B)}$; it would then be possible to measure the noise power due to the input circuit of the receiver with an accuracy of $T_R\sqrt{(b/B)}$.

If now the aerial is connected to the receiver, the output reading will fluctuate in a random way about a new mean value corresponding to an input of $T_R + T_A$. The increase in the output reading will be detectable if it exceeds the fluctuations; the aerial power will therefore be detectable if $T_A > T_R\sqrt{(b/B)}$. It is thus possible to detect a random power from the aerial which is considerably smaller than that inherent in the receiving amplifier by using an amplifier of large bandwidth followed by an output indicator of small bandwidth. The ultimate sensitivity which may be achieved in practice will be discussed later, but it should be noted that there are certain fundamental limitations which may determine the smallest power which can be detected. In the first place it may be important to examine the way in which the intensity of a source varies with the frequency, an experiment which requires the use of a comparatively narrow input bandwidth; it will be seen later that certain measurements with high resolving power also require a limited bandwidth due to the different phases of the interference patterns produced by the various components passed by a wide-band receiver. In the second place the use of a very narrow bandwidth in the output indicator may be prevented by the necessity for observing comparatively rapid changes of the received intensity; such changes might be due to variations of the incident radiation, or variations in the absorbed power which are caused by the rotation of the aerial system. In practice values of B/b of 10^5–10^7 are frequently used; it should then be possible to detect an aerial power of about one-thousandth of the noise power due to the receiver itself.

Whilst theoretically possible to detect small powers from the aerial in this way, the design of a practical system presents considerable difficulty, particularly since it is usually necessary for the apparatus to operate for long periods without calibration, and with the output automatically recorded. The addition of a narrow-band filter to the output of a conventional receiver may make it possible to reduce the residual fluctuations to a value which corresponds to the required sensitivity, but the mean value of the output is proportional to the receiver gain, and thus a change in the gain of the order of 0·1% will produce a change in the output reading corresponding to the required limit of detection. Variations of the noise power produced in the receiver itself (due, for example, to deterioration of the first valve) would also cause a proportional change in the output reading. The maintenance of the gain or noise power of a high-gain amplifier to this order of accuracy is not practicable without periodic calibration.

These difficulties have been partly overcome by Dicke (1946) in a method which periodically replaces the aerial by a resistance at room temperature; the period of the switching is sufficiently short for the constants of the receiver to remain unaltered between successive positions of the switch. If the aerial and resistance present equal impedances to the input of the receiver, the output of the receiver will contain a component at the switching frequency whose amplitude is proportional to the gain of the receiver and to the difference between the equivalent aerial temperature and room temperature. A two-fold increase

of the output reading will therefore be produced by a doubling of the receiver gain; in the simple system a two-fold increase of the output reading might be caused by a variation of the gain of 0·1%.

The output does, however, still depend on the linearity of the receiver and on the law of the detector and, in general, calibration of the apparatus is necessary. Unless such calibration is carried out by means of a standard source of noise power it is necessary to know the precise frequency response of the receiver.

An alternative method of measuring small noise powers has been developed (Ryle and Vonberg 1948) which is independent of the gain and frequency response of the receiver and does not involve a knowledge of the law of the detector. In addition, it is capable of operating over a very wide range of input powers without the necessity for calibrated switching of the gain.

In this system, which is shown diagrammatically in Figure 2, a local source of noise power is continuously and automatically adjusted so as to be in equilibrium with the aerial. The receiver is switched alternately between the

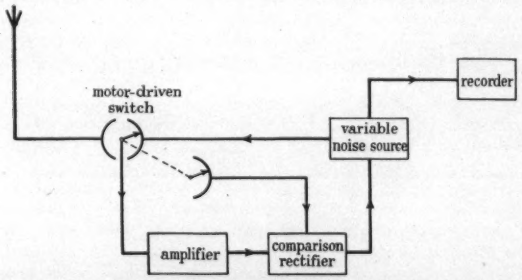


Figure 2. System for the continuous measurement of small 'noise' powers.
(*Reproduced by kind permission of the Royal Society.*)

aerial and the local source of noise power; if the impedances of aerial and noise source are equal, the powers flowing into the receiver in the two positions will be equal when the effective temperature of the noise source is equal to that of the aerial. This condition is independent of the impedance and noise level of the receiver itself. If the powers are not equal, the output of the receiver will contain a component at the switching frequency whose amplitude and phase are a measure of the unbalance between aerial and noise source. By comparing this component with a reference waveform derived from the switch-motor shaft it is possible to derive a D.C. voltage proportional to the 'error' in the output of the noise source, and to use this voltage to control the noise source towards equality. The final adjustment does not depend on the gain or frequency response of the receiver, or on the linearity of the detector.

A diode operating under temperature-limited conditions has been used as a source of noise power; if the anode current, I, from such a valve is allowed to flow through a resistance R (chosen to present the same impedance to the receiver as the aerial) the available noise power in a bandwidth δf is given by

$$\tfrac{1}{2}eIR\delta f + kT_0\delta f,$$

where e is the electronic charge and k is Boltzmann's constant. (The second term represents the Johnson noise in the resistance.) This power is the same as would be obtained if the resistance were heated to a temperature T_0+T_D, where $T_D=eIR/2k$.

The equivalent aerial temperature may thus be determined from the mean anode current of the diode, in terms of a single parameter, the resistance used in the diode anode circuit.

The operating range of the instrument may be extended by inserting attenuating cables between the noise source and the change-over switch. It can be shown that a resistance at a temperature T connected through a matched transmission line having a power attenuation constant α will produce the same available power as a resistance at a temperature $\alpha T+(1-\alpha)T_0$, where T_0 is the temperature of the transmission line.

If, therefore, an attenuating cable is used in the output of a diode noise source, the effective temperature is given by $\alpha(T_0+T_D)+(1-\alpha)T_0=T_0+\alpha T_D$. The sensitivity of the equipment is therefore decreased by a factor α, and the zero level still corresponds to room temperature.

If an aerial, whose equivalent temperature is T_A, is connected to the change-over switch by a cable of attenuation constant β, the balance conditions will be given by

$$\beta T_A+(1-\beta)T_0=T_0+\alpha T_D$$

or

$$T_A=T_0+(\alpha/\beta)T_D.$$

The diode anode current may thus be used to provide a direct record of aerial power, in which the zero is always room temperature, and whose scale depends on the resistance in the anode circuit of the diode, R, and the ratio of the attenuation constants of the diode and aerial cables.

At the shorter wavelengths the aerial temperature is found to be less than the room temperature. It is possible to displace the zero level of the instrument by a small intentional mismatch of the aerial, although under these conditions the zero level is somewhat dependent on the noise level of the receiver. Whilst such a dependence is not of significance when interference-type aerials are used (see §5), it is often undesirable. An alternative method which has been used at a wavelength of 60 cm. interchanges the resistance R and the aerial; the latter now acts as the load resistance for the diode. With this arrangement the apparatus may be used to record aerial temperatures below room temperature, the relation being $T_A=T_0-(\alpha/\beta)T_D$ (Stanier, unpublished).

Design details for the automatic balancing system and a full analysis of the effects of mismatch of the aerial impedance and of the response of the system to a rapidly varying aerial temperature are given by Machin, Ryle and Vonberg (in preparation).

§5. THE OBSERVATION OF A SMALL SOURCE AGAINST THE BACKGROUND RADIATION OF THE GALAXY

In addition to the problem of measuring the small random powers which are available from the aerial system, a further difficulty arises when measuring the intensity from a source of comparatively small angular diameter, such as the sun; this difficulty is due to the masking effect of the radiation from the galaxy. Over most of the range of wavelengths which have been used in Radio Astronomy

the solid angle for reception, Ω, is considerably larger than the solid angle subtended by the sun, ω. If the emission from the sun corresponds to black-body radiation at a temperature T_s, and if the emission from the galaxy is assumed to be constant over the solid angle for reception, and to have an intensity corresponding to a temperature T_g, the ratio between the aerial power due to the sun and that due to the galaxy is

$$\frac{\omega T_s}{\Omega T_g}.$$

It may therefore be impossible to detect the radiation from the sun unless $T_s \gg T_g$. As an example, the radiation at a wavelength of 4 metres from the undisturbed sun corresponds to a temperature T_s of about a million degrees, whilst that from near the galactic equator corresponds to a temperature T_g of about 5,000 degrees. If an aerial having a solid angle for reception, Ω, of one steradian (corresponding to a radiating area of about 20 m^2) were used, the component due to the sun would only be about 1% of that due to the galaxy. In order to produce a solar component equal to that from the galaxy it would be necessary to use an aerial having an area of about 2,000 m^2.

The measurement of the intensity of the radiation from the undisturbed sun and, as will be seen later, the detection of discrete sources of radiation in the galaxy, are thus likely to be limited by the background radiation from the galaxy, and a more detailed analysis of the limit of detection will now be made. In this analysis it will be assumed that the recording receiver itself has an adequate limiting sensitivity.

Two main types of aerial system have been developed for discriminating against the galactic background: the first involves the use of a large radiating area to produce a 'pencil' beam (Ω small); the second makes use of the interference pattern produced by a pair of spaced aerials to provide a considerably greater discrimination with an aerial system of given total area, or conversely, to obtain a given degree of discrimination with a much simpler aerial system. The spaced-aerial system is analogous to Michelson's interferometer, and many of the advantages of the optical system will be seen to apply equally to the radio version.

The two methods will first be examined in relation to the measurement of solar radiation; the more difficult problem of detecting discrete sources in the galaxy will then be discussed.

(*a*) *The 'Pencil' Beam*

The technique of the 'pencil' beam, which is familiar in optical astronomy, requires the use of an aerial system having the greatest possible aperture. As was shown earlier, an aerial having an effective radiating area A will produce a radiation pattern whose effective solid angle Ω is given by $\Omega = \lambda^2/A$, where λ is the wavelength. (It is also worth noting that the power picked up by directing such an aerial at a remote source, compared to that received by an isotropic system, is given by $G = 4\pi A/\lambda^2$. This characteristic of an aerial is clearly of importance when the limit of detection is determined by the ultimate sensitivity of the recording receiver; when the limit is set by the background radiation from the galaxy it is only necessary to consider the solid angle for reception, Ω.)

The measurement of solar radiation using a 'pencil' beam involves, basically, a comparison of the received power when the aerial is directed at the sun, with

that when the aerial is turned away from the sun, through an angle sufficiently large for the contribution of the sun to be negligible. Ideally this angle would be related directly to $\sqrt{\Omega}$, but in practice the presence of secondary maxima in the radiation pattern may make it necessary to turn the aerial through a considerably larger angle.

On the assumption that the power produced in the aerial by the galactic background is identical in the two positions, the intensity of the solar component can then be determined. Even in the absence of discrete galactic sources of very small angular diameter, it is clear that this method is limited by the angular variations of the galactic background; if the angular structure of the galaxy is comparable with the width of the radiation pattern, it is clearly impossible to measure the radiation from the sun unless it produces a power in the aerial of the same order as that due to the galactic background.

At the shorter wavelengths (less than about 50 cm.) the ease with which a small beam width may be achieved with a structure of reasonable size, together with the fact that the intensity of the galactic radiation is relatively small compared with that from the sun, make the 'pencil' beam very suitable for observing solar radiation. At the longer wavelengths it becomes increasingly difficult to construct aerials of sufficiently large dimensions to produce a radiation pattern whose angular width is small compared with the general structure of the galaxy; under these circumstances, and with the relatively much greater intensity of the galactic radiation, it is difficult to associate the small variations of the recorded power, as the aerial is moved through an angle equal to its beam width, with the radiation from an individual weak source, such as the undisturbed sun.

It is difficult to state the minimum aerial aperture necessary to measure a given flux of solar radiation, since it depends not only on the intensity of the galactic radiation, but also on its rate of change with angle in the direction of the sun. For a typical case of observing radiation from the undisturbed sun on a wavelength of 4 metres ($T_s = 10^6$ deg. K., $T_g = 5000°$ K.) it would probably be sufficient to use an aerial system having an aperture of about 200 m^2. The solar component from such an aerial would then represent about 10% of the galactic component, whereas the change in background due to the rotation of the aerial through its beam width (about 15°) should represent a somewhat smaller variation. From such an example, however, it is clear that the accurate measurement of solar radiation at the longer wavelengths over long periods (as the position of the sun relative to the galaxy is changing) requires the use of an aerial system sufficiently large for the solar component to be of the same order as the total galactic component in the aerial.

(*b*) '*Interference*' *Method*

In this method the interference between two aerials is used to produce a radiation pattern in which a number of interference maxima and minima occur within the normal primary radiation pattern of each aerial. Two types of interferometer have been used: one is based on the interference between an aerial system mounted on a high cliff with the image produced by reflection at the surface of the sea (McCready, Pawsey and Payne-Scott 1947). The other is based on the interference between two similar aerials spaced laterally in a direction normal to the wave front (Ryle and Vonberg 1946, 1948).

If an interference pattern of this type is moved across the sky (e.g. by the earth's rotation), the passage of successive maxima across the sun will give rise

to a periodically varying signal on the record. If the angular separation of the maxima of the interference pattern is small compared with the angular structure of the galaxy, no similar periodic variation will be produced by the galactic radiation; the component due to the galactic background will produce a steady trace, or one which varies slowly, as the primary radiation pattern of each aerial (the envelope of the interference pattern) is moved across the galactic plane. In order that the varying component on the record may be used to indicate the intensity of solar radiation independently of its distribution across the solar disc, it is important that the intensity due to the sun at the minima shall be very

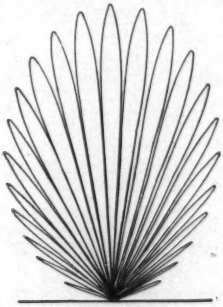


Figure 3. Radiation pattern produced by two aerials of aperture 1λ, spaced 10λ apart.
(Reproduced by kind permission of the Royal Society.)

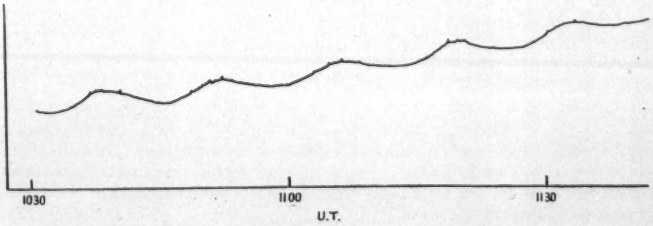


Figure 4. Record obtained on a wavelength of 3·7 m. with the spaced-aerial system, showing radiation from the undisturbed sun and from the galaxy.

small. (This requirement is analogous to the visibility of the fringes in Michelson's optical interferometer being complete.) Since radiation may be emitted from any point inside the solar disc (and, as will be seen later, even from the corona at the limb) it is important to choose an aerial spacing which will produce maxima separated by an angle which is several times that of the sun's diameter. Figure 3 shows the interference pattern produced by two aerials of small directivity at a spacing of 10 wavelengths (Ryle and Vonberg 1946); this system has been used extensively to record solar radiation on a number of wavelengths, and a typical record obtained on 3·7 m. is shown in Figure 4.

The adoption of an interference system makes it possible to obtain a discrimination against the galactic background, which is equal to that of a 'pencil' beam having an angular width a few times the angular diameter of the sun; the production of such a 'pencil' beam would require an aerial aperture about twice the spacing between the two aerials of the interference method. In the interference system the discrimination is not critically dependent on the aperture of the individual aerials, and very simple aerials are frequently used. (At the shorter wavelengths, where the detection of solar radiation is also limited by the sensitivity of the recording receiver, it is important to use aerials of comparatively large area in order to increase the available aerial power, but at longer wavelengths the limit of detection is usually set entirely by the masking effect of the galactic background.) Recent experiments at Cambridge (Machin, unpublished), using an interference method with a somewhat modified receiving system, have shown that it is possible to obtain regular observations of solar radiation on a wavelength of 6·7 metres, when each aerial has an extremely small directivity ($G=8$); in this case the aerial power due to the undisturbed sun represents only about 0·1% of that due to the galaxy.

The ability to use aerials of small directivity is a great advantage, particularly at the longer wavelengths, not only because of their simpler construction, but because it makes it possible to observe the sun for a considerable period each day without the need for continuously rotating the very large aerial structure which would be necessary in a 'pencil' beam method.

A further advantage of considerable practical importance is related to the absolute measurement of solar radiation. As was shown in §4, small changes of the impedance of the aerial system from day to day can produce a displacement of the zero of the record; such a displacement would affect measurements made with the 'pencil' beam method, but would not alter the amplitude deduced in the interference method.

On the other hand, the use of an interference method results in the loss of observation for short periods when the sun is near the minima of the interference pattern, and short-lived variations of the intensity may be missed. This has not been found to be a serious limitation in practice.

The application of interference methods to determine the size and position of sources will be considered in §5 (*d*).

(*c*) *The Detection of Discrete Sources in the Galaxy*

So far it has been assumed that the background radiation from the galaxy is uniform over angles which are small compared with the general structure of the galaxy. It is now known that there are discrete sources of very small angular diameter which contribute at least an appreciable fraction of the total galactic radiation; it is possible that in fact the whole of the radiation is due to a large number of these sources distributed throughout the galaxy, and that only the most intense of them have been resolved with existing apparatus. This possibility will be discussed later in §10; the problems of detecting the sources will be discussed here.

It will be obvious that the presence of discrete sources in the galaxy might have a serious effect on the measurement of solar radiation, since the methods which have been proposed for discriminating against the galactic radiation were based on the assumption that the angular distribution of galactic radiation was constant over angles comparable with that subtended by the sun. In practice

it is found that only two of the discrete sources produce an intensity at the earth which is comparable with that from the undisturbed sun, and neither of these sources is ever located in a direction near to that of the sun; for the purpose of solar measurements it is therefore legitimate to regard the galaxy as a diffuse source whose angular variation of intensity is comparable with the structure of the galaxy.

It is now necessary to apply the previous arguments to the detection of discrete galactic sources; the angular diameters of these sources are less than that of the sun, but their intensities relative to that of the background radiation may be very much smaller, and it is important to examine the ultimate limit of detection of the different methods.

It is clear that with the 'pencil' beam technique, the construction of larger and larger aerials will produce a progressive improvement in discriminating against the 'diffuse' radiation from the galaxy; if the beam width were small compared with the angular structure of the galaxy, detection of discrete sources would only be limited by confusion with other discrete sources. An aerial having sufficiently large dimensions to produce a radiation pattern small compared with the structure of the galaxy has not yet been built for any wavelength at which galactic radiation has been studied; if such an aerial were built, however, the variation of received power which would occur as it was moved across a discrete source would correspond to its own radiation pattern. The maximum number of discrete sources which could be identified in this way is clearly related to the solid angle for reception, and for a number of uniformly distributed sources of equal intensity would be of the order of $G/4$, where the power gain $G = 4\pi A/\lambda^2 = 4\pi/\Omega$.

For all aerial systems which have so far been used to observe discrete galactic sources, however, the power gain has been far too small for the variations due to the structure of the galaxy to be disregarded; the variations of received power which occur when the aerial is moved are therefore largely due to variations in the 'diffuse' background, and the number of the variations which can be reliably identified as discrete sources will be very much fewer.

The use of an interference system, on the other hand, makes it possible to eliminate all variations of intensity except those due to sources of very small angular diameter without the use of complicated aerials of large area. The discrimination of such a system against the variations of the 'diffuse' background corresponds to that of a 'pencil' beam aerial whose aperture is twice the *spacing* between the aerials. It is therefore possible to obtain any required degree of discrimination against the background radiation by using two aerials of relatively small directivity spaced a considerable distance apart. Observations have been made at Cambridge on a wavelength of 3·7 m. using two aerials each of which has a solid angle for reception of about 0·05 steradians (corresponding to a power gain of about 250) with a spacing of 110 wavelengths (about 400 m.); the aerial system and radiation pattern are shown diagrammatically in Figure 13. With this system it has been possible to identify discrete sources whose intensity is only 10^{-5} of that of the galactic background. (This figure corresponds to the detection of a varying component of aerial power which is 10^{-3} of the total aerial power.)

The maximum number of discrete sources which can be detected with an interference system is therefore determined solely by the overlapping of the interference patterns produced by adjacent sources (in the absence of limitations

based on the sensitivity of detection). It should therefore be possible to detect about $G/4$ sources, where G is the power gain of each aerial. (It will be shown later in § 8 that some discrimination is possible even if the patterns of two sources do overlap, since sources at different declinations give rise to patterns having different periods.)

The total number of sources which could be detected by the 'pencil' beam system and the interference system having equal radiating areas would therefore be approximately equal for very large radiating areas, where the radiation pattern of the 'pencil' beam is sufficiently small to discriminate against the 'diffuse' background. No aerial of sufficiently large dimensions has yet been constructed for observing galactic sources, however, and for aerials of smaller dimensions the interference system has considerable advantages, since the confusion caused by variations of the 'diffuse' background radiation can always be eliminated, and the limitations set by confusion with other discrete sources should make it possible to detect about $G/4$ uniformly distributed sources of equal intensity, for all values of G, the power gain of each aerial. The performance of a 'pencil' beam system, however, falls off rapidly as G is decreased, owing to the difficulty of distinguishing between the variations caused by the structure of the galaxy and those due to genuine discrete sources.

In addition to its advantages in detecting discrete sources in the galaxy, the interference system has very important applications for the accurate location of the sources; the accuracy with which the position may be found is the same as that for a 'pencil' beam system having an aperture about twice the spacing between the aerials of the interference system. This application will be described in §§ 5 (*d*) and 8 (*b*).

In the above analysis it has been assumed that the detection of the signals presents no fundamental limitation; in many cases it is important to consider the overall limit of sensitivity in addition to problems of discrimination. It is necessary to detect a small additional signal which exhibits variations as a 'pencil' beam or interference pattern is moved across a discrete source. At wavelengths shorter than about 3 metres the limit of detection will be determined by the 'noise' generated in the receiver, but at longer wavelengths the statistical fluctuations of the received background radiation from the galaxy constitute a 'noise level' which can only be reduced by statistical averaging; the time available for this averaging is clearly related to the rate at which the contribution from the discrete source will change as the radiation pattern of the aerial system is moved. For all systems of high resolving power which involve the use of aerials fixed to the earth there is thus a fundamental limit to the output bandwidth (and therefore the sensitivity) of the recording receiver, which is set by the angular resolving power of the aerial system, and the rotational velocity of the earth.

This consideration is of some importance when accurate position finding of weak sources is attempted.

(*d*) *The Measurement of the Diameter and Position of a Source*

As will be seen later, the determination of the size and position of the sources of intense solar radiation is of great importance, and the extension of the interference technique to these measurements will now be discussed. The application of the method to the location of discrete galactic sources will be left until later (§ 8 (*b*)).

As has already been mentioned, a source whose angular dimensions are very small compared with the angular separation of the maxima of the interference pattern will produce a record whose minima fall virtually to zero. If, however, the source subtends an angle comparable with the separation of the maxima, the minima will be incomplete, and a measurement of the maximum to minimum ratio enables the source diameter to be determined. The method is analogous to Michelson's method for determining the angular diameters of stars, the maximum to minimum ratio on the record corresponding to the 'visibility' of the optical fringes. Both the spaced aerial method, and the cliff method, which is analogous to a 'Lloyd's mirror' interferometer, have been used for determining the size of the sources of the enhanced solar radiation associated with sunspots.

By observing the precise moment at which the central maximum of the interference pattern occurs, it is also possible to obtain a very accurate determination of the direction of the source. Both types of interferometer have again been used, and with spacings of 100–200 wavelengths it has been possible to locate the active regions of the sun's disc with an accuracy of 1–2 minutes of arc. With aerial spacings of this order of magnitude, the separation of successive maxima is comparable with the angular diameter of the sun, and no difficulties arise in connection with the identification of the central maximum.

Similar methods have been used to locate discrete galactic sources; it now becomes necessary, however, to have some independent method of identifying the central maximum on the record. Two methods have been used: the first makes use of a restricted primary radiation pattern, so that the central maximum can be identified by its amplitude; the second involves the use of two different aerial spacings to produce records of different periodicity which have the same phase at the central maximum. The application of these methods and the determination of the declination or North–South coordinate of the position will be described in § 8(*b*).

In many applications of the interference method it is important that observations should be restricted to the central fringes; in high-order fringes the phase of the interference pattern depends on the wavelength, and since the band of wavelengths accepted by the receiver is finite (owing to requirements of limiting sensitivity), the observed maximum to minimum intensity on the record will be reduced. Any observations of source diameter which were made at an appreciable angle to the central region would therefore be misleading, since the reduced ratio of maximum to minimum intensity would be interpreted as an apparent increase of diameter. This effect is analogous to the formation of white light fringes near the centre of an optical interference pattern, with coloured fringes at wide angles; the 'visibility' of the fringes as recorded on a detector sensitive to a wide band of wavelengths would be at a maximum at the centre but would decrease as the angle from the centre was increased.

The effect may be of considerable importance when wide bandwidths are used to obtain the maximum possible sensitivity; for example, observations of the sun on a wavelength of 1·7 m. with a bandwidth of 2 Mc/s. ($\delta f/f \simeq 0{\cdot}01$) and an aerial spacing of 140 wavelengths had to be restricted to within 40 minutes of local noon (Ryle and Vonberg 1948).

An extremely interesting application of the spaced aerial interferometer has been used to determine the distribution of the intensity emitted by different parts of the solar disc in the absence of sunspots. In this experiment the

amplitudes of the records obtained with a number of different aerial spacings are measured; each of these amplitudes corresponds to one term of the Fourier transform of the initial angular distribution of the intensity in the source, and by assuming circular symmetry it is possible to deduce the actual distribution across the solar disc. Details of this application will be given in § 7 (*a*).

§ 6. THE MEASUREMENT OF POLARIZATION

The emission of radiation by a completely random mechanism, such as the emission of light by a heated solid, produces a field in which the two mutually perpendicular components exhibit no phase coherence. Observations of the radio waves emitted by the sun have indicated that occasionally the radiation is not randomly-polarized, and that a circularly-polarized component can be detected.

It is therefore important to be able to carry out a complete analysis of a complex wave emitted by the sun or by sources in the galaxy. The wave might consist of a mixture of elliptically-polarized and unpolarized radiation, and it might therefore be necessary to deduce the relative intensities of the random, circular and plane components. The methods of analysis which have been used are analogous to those familiar in the analysis of polarized light; a linearly-polarized aerial takes the place of the Nicol prism analyser, whilst quarter-wave plates are used in both the radio and optical systems, although more usually circularly-polarized components are detected by means of two mutually perpendicular linearly-polarized aerials having a displacement of a quarter-wave. So far no plane-polarized components have been detected, and the majority of polarization measurements have therefore been concerned with the determination of the intensity and polarity of the circularly-polarized component relative to the randomly-polarized component.

Two cases must be considered: (*a*) where the intensity of the radiation from the source is large compared with the total intensity due to the background; and (*b*) where the source produces an intensity comparable with or smaller than the background radiation.

In the first case it is clear that the presence of a plane-polarized component can be determined by rotating a linearly-polarized aerial system about the wave normal and observing the variation of the recorded intensity with the position of the aerial; a randomly-polarized or circularly-polarized wave will produce a recorded intensity which is independent of the aerial position, while the difference between the maximum and minimum recorded intensities will enable any plane-polarized component to be measured.

The presence of circularly-polarized radiation may be determined by the use of two adjacent linearly-polarized aerials whose planes of polarization are normal to one another. If the outputs of the two aerials are combined in such a way that an extra quarter-wave is introduced in one or other of the paths, the system will be sensitive to a right- or left-handed polarized wave, but will produce zero output for a wave of the opposite polarization. A randomly- or plane-polarized wave will produce an equal power in the two positions, having a value which is half that produced by the appropriate circularly-polarized wave of equal intensity. By observing the recorded intensity in the two alternative conditions it is therefore possible to determine the intensity and polarity of the circularly-polarized component and the intensity of any plane or randomly-polarized

components. The presence of two equal circularly-polarized waves of opposite polarity cannot, of course, be distinguished from randomly-polarized radiation by this method.

The extra quarter-wave may be introduced into one aerial either by displacing the two aerials by a quarter of a wavelength along the wave normal, or by inserting an extra quarter-wave of transmission line in one of the aerial connectors. Unless the aerials are accurately matched to the transmission line over the band of frequencies accepted by the receiver, however, the insertion of an extra quarter-wave of transmission line may result in the two aerials presenting unequal impedances at the junction point. Under these conditions the system will not provide complete rejection of the unwanted circularly-polarized component; the physical displacement of the aerials by a quarter-wavelength, with transmission lines of equal length, or differing by half a wavelength (which introduces no impedance transformation), is therefore a more satisfactory arrangement.

At the shorter wavelengths it becomes possible to construct a quarter-wave plate operating on the principle of a waveguide; by mounting it in front of the aerial system with its axis at an angle of 45° to that of the aerial the system will become sensitive to one or other of the circularly-polarized components. Observations of the polarization of solar radiation on a wavelength of 10·7 cm. have been made in this way by Covington (1949).

If the radiation from the source under investigation is masked by an equal or greater intensity of randomly-polarized radiation from the galactic background, it is clearly difficult to determine the degree of polarization by the methods described, particularly since small changes of polar diagram or impedance caused by the rotation or switching of the aerials are likely to produce small displacements of the recorded background intensity. An alternative method of measurement has therefore been developed which avoids any movement or switching of the aerials. In this system two linearly-polarized aerials are spaced on an East–West line and the periodic variations of intensity caused by the earth's rotation are used to distinguish the radiation due to the wanted source from that due to the background. In order to detect the presence of a linearly-polarized component it is necessary to observe the amplitude of the interference pattern obtained with the aerials parallel and rotated to different positions about the wave normal.

If the aerials are now mounted in mutually perpendicular directions (see Figure 5) the system may be used for determining the intensity of the circularly-polarized component. The incidence of a randomly-polarized wave will produce components in the two aerials which have no phase coherence, and variations of the path difference to the two aerials as the earth rotates will cause no variation of the recorded intensity. The presence of randomly-polarized radiation is therefore equivalent to an increase in the background radiation. The incidence of a circularly-polarized wave, on the other hand, will produce coherent components in the two aerials, and a variation of the path difference to the two aerials will give rise to a periodically varying trace on the record; the amplitude of this trace for a fully-polarized wave will be equal to that obtained with the aerials parallel, but it will be displaced in time by a quarter of a period, in a direction which depends on the polarity of the wave. By comparing the amplitudes obtained with the aerials parallel and mutually perpendicular, it is therefore possible to determine

the polarity and degree of polarization in the incident radiation. The type of record obtained for incompletely polarized solar radiation is shown in Figure 6. It should be noted that a plane-polarized wave having a plane of polarization intermediate between those of the two aerials will also produce a periodically varying signal, but in this case it will have the same phase as that obtained with the aerials parallel.

The spaced-aerial interferometer has been used to determine the polarization of weak radiation from the sun (Ryle and Vonberg 1948) and from discrete sources of radiation in the galaxy (Ryle and Smith 1948). While having the great

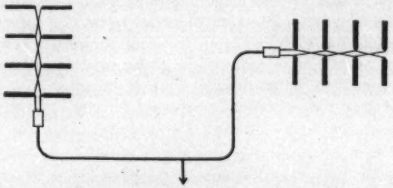


Figure 5. Interference polarimeter, viewed along the wave-normal.
(*Reproduced by kind permission of the Royal Society.*)

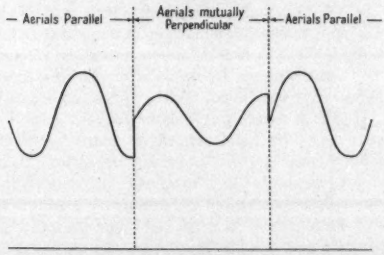


Figure 6. Diagram to indicate the type of record obtained with the interference polarimeter.

advantage of allowing polarization measurements in the presence of a large unpolarized background, it is less convenient for observing the polarization of the short-lived 'bursts' of radiation associated with sunspot activity. Under such conditions, however, the intensity is usually sufficiently great to allow the use of the direct method.

§7. EXPERIMENTAL OBSERVATIONS OF SOLAR RADIATION

Two comprehensive survey papers covering the emission of radio waves from the sun have recently been written (Hey 1949, Pawsey 1950) which give a detailed historical survey of the experimental observations. No attempt will therefore be made in this section to provide a complete picture of the experiments; the aim will be to give a summary of the outstanding conclusions of these observations, without regard to their historical order.

It is convenient to divide the wavelengths available for solar observation into two ranges: (*a*) those shorter than 1 m. and extending down to 1 cm.; (*b*) those longer than 1 m. and extending at the present time to about 7 m., although certain short-lived phenomena have been observed at wavelengths as long as 20 m. In the region of wavelengths shorter than 1 m. the background radiation from the galaxy is weak, and since aerials of considerable directivity can also easily be built, the radiation from the undisturbed sun has been widely studied. At the longer wavelengths there is less information on the radiation from the undisturbed sun, but many observations have been made on the greatly enhanced radiation which, at these longer wavelengths, is associated with sunspots.

A diagram showing, as a function of wavelength, the approximate relative powers which would be produced in an aerial of constant physical dimensions is shown in Figure 7. The powers produced by galactic and solar radiation in this diagram are deduced for an aerial having an effective area of 10 m². For aerials of smaller area, the contribution of the sun relative to that of the galaxy

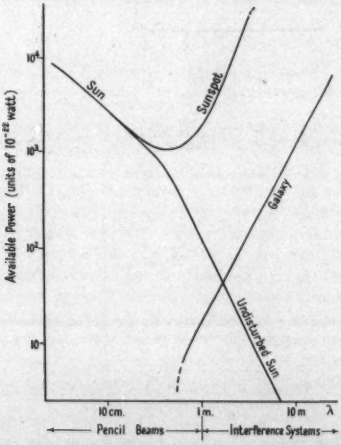


Figure 7. Diagram to show, as a function of wavelength, the variation of power available from an aerial (of area 10 m²), due to the sun and the galaxy.

would be correspondingly reduced. It should, of course, be remembered that the construction of an aerial of given physical dimensions is relatively much simpler at a longer wavelength; at a wavelength of 10 cm. aerials having an area of 5 m² have been used, whilst areas of over 100 m² have been used at wavelengths of a few metres.

(*a*) *Observations on Wavelengths shorter than* 1 *m.*

No comprehensive study of solar radiation in the band of wavelengths from 1 cm. to 1 m. has yet been made, but detailed measurements have been made in a number of narrow bands at about 1, 3, 10 and 50 cm. Some observations have also been made at 25 cm. The pioneer work was done by Southworth

(1945), who demonstrated that solar radiation could be detected on wavelengths of 1, 3 and 10 cm., and who deduced that the sun's emission at these wavelengths corresponded to a temperature not greatly different from that determined by visual measurements (6,000° K.).

Subsequent observations at 1 cm. by Dicke and Beringer (1946) during a partial eclipse showed that the intensity corresponded to a source having the same size as the visual disc, and at a temperature of about 10,000° K. Similar observations on a wavelength of 3 cm. were made by Sander (1947), while Schulkin *et al.* (1948) observed that during a chromospheric flare the intensity of the 3 cm. radiation showed a sudden and brief enhancement.

A detailed study of the 10 cm. radiation from the sun has been made by Covington (1947, 1948, 1949). He has shown that, in the absence of sunspots, the sun emits radiation corresponding to a temperature of about 80,000° K., and by observations made during a partial eclipse was able to show that small areas in the neighbourhood of sunspots radiated with very much greater intensity (corresponding to a temperature of $1{\cdot}5 \times 10^6$ deg. K.). By regular observations over a period of eight months Covington demonstrated that the intensity of solar radiation at a wavelength of 10 cm. showed good correlation with the sunspot number. The enhanced radiation associated with sunspots was found to be circularly-polarized. Small, sudden increases of intensity were also sometimes observed at times of chromospheric flares.

Observations by McCready, Pawsey and Payne-Scott (1947) on a wavelength of 25 cm. indicated that the sun normally emitted as if it had a temperature of about 2×10^5 deg. K., while Lehany and Yabsley (1948) showed that the intensity on this wavelength was also correlated with sunspot area.

More extensive data are available for wavelengths of 50–60 cm. Reber (1946, 1948 a) showed that the intensity exhibited small (15%) variations from day to day, which appeared to be associated with the presence of sunspots, whilst occasional short-lived increases of intensity were observed. These results were confirmed by the work of Houtgast and Laffineur (1948). Extended observations over a period of nine months at Cambridge have shown that in the absence of sunspots the intensity corresponds to the radiation from a 'black' source subtending an angle of $\frac{1}{2}°$ and at a temperature of $5{\cdot}5 \times 10^5$ deg. K. (Stanier 1950). The observations have also shown variations of mean daily intensity of up to 100% when large sunspots are crossing the disc, whilst occasional 'outbursts' greater than ten times the intensity from the undisturbed sun have been observed which appear to be closely related to chromospheric flares. The enhanced radiation associated with sunspots usually exhibits slight circular polarization (about 20%), and there is evidence that the radiation associated with the 'outbursts' is more fully polarized. The radiation from the undisturbed sun has been found to be unpolarized.

In the above description it has been assumed that the radiation from the undisturbed sun is emitted uniformly from a disc having the same diameter as the photosphere (about $\frac{1}{2}°$). As will be seen later (§ 9(*c*)), some theories of the emission of radio waves from the sun predict that at wavelengths of 50–60 cm. (and also at shorter wavelengths) pronounced 'limb-brightening' will occur; the radiation on these theories will be emitted mainly from a narrow ring round the circumference of the sun. Both from the point of view of proving the theory, and because the smaller solid angle of such a source would involve the occurrence

of still higher effective temperatures to account for the observed intensity, it is important to attempt an experimental investigation of the distribution of intensity of the 50 cm. radiation from the sun. Two methods have been used: the first, due to Laffineur, Michard, Steinberg and Zisler (1949), involved the observation of the variation of intensity during a partial eclipse. The results obtained at wavelengths of 25 cm. and 55 cm. suggested that at both these wavelengths emission from a 'ring' source was unlikely.

A similar experiment has recently been carried out by Christiansen, Yabsley and Mills (1949) on a wavelength of 50 cm.; although they were able to show that a considerable fraction of the total radiation from the undisturbed sun was emitted by regions near or outside the limb, they were not able to distinguish conclusively between a source showing limb-brightening and a source of uniform 'brightness' having a radius of $1{\cdot}3R_0$ (where R_0 is the radius of the photosphere).

Observations based on the variation of the received intensity during an eclipse are difficult since (except for eclipses which are almost complete) the shape of the observed variation of intensity is not very sensitive to the degree of limb-brightening.

The second method is due to Stanier (1950); this method allows the distribution across the solar disc to be measured directly, and therefore allows observations to be made at times when the sunspot activity is small. The previous method may be complicated by the presence of sunspots at the time of the eclipse. In Stanier's method a spaced-aerial interferometer is used, and the amplitudes of the periodic variations on the record which are obtained with a number of different aerial spacings are determined.

The receptivity (power) of a spaced-aerial interferometer in a direction θ from the central maximum is given by $2F(\theta)[1+\cos(2\pi d\sin\theta/\lambda)]$, where $F(\theta)$ represents the receptivity of each aerial and d is the spacing between them. For small angles of θ the variations due to the primary radiation pattern, $F(\theta)$, can be neglected, and the receptivity is given by $a\{1+\cos(2\pi d/\lambda)\theta\}$.

Now consider the power received from an extended source whose distribution of intensity with θ may be represented by a narrow strip source parallel to the line joining the aerials, and having a distribution $f(\theta)$. The power available from the aerial when the centre of the source is at an angle θ_1 from the central maximum of the interference pattern may be represented by

$$a\int[1+\cos\{(2\pi d/\lambda)(\theta+\theta_1)\}]f(\theta)\,d\theta,$$

where the integration extends over the whole source. This can be re-written

$$a\int f(\theta)\,d\theta+a\cos\left(\frac{2\pi d}{\lambda}\theta_1\right)\int\cos\left(\frac{2\pi d}{\lambda}\theta\right)f(\theta)\,d\theta$$

$$-a\sin\left(\frac{2\pi d}{\lambda}\theta_1\right)\int\sin\left(\frac{2\pi d}{\lambda}\theta\right)f(\theta)\,d\theta.$$

Now consider the variation of the aerial power which will be caused as the interference pattern is moved across the source by the rotation of the earth.

Disregarding the constant component of the aerial power, it can be seen that the variable component is due to the two terms of the Fourier transform of $f(\theta)$ which have a periodicity in angle of d/λ. On the assumption that the function $f(\theta)$

is symmetrical, the term in $\sin\{(2\pi d/\lambda)\theta\}$ is zero, and the amplitude of the observed variation of received power is given by

$$2a\int \cos\left(\frac{2\pi d}{\lambda}\theta\right) f(\theta)\, d\theta.$$

It can therefore be seen that each observation of the source with a spaced-aerial interferometer will provide one term of the Fourier transform of the distribution of intensity; by taking a number of different observations with different spacings of the aerials, d, it is therefore possible to deduce the original distribution $f(\theta)$.

This process is exactly analogous to the method used by Michelson for determining the fine-structure of spectral lines; by observing the 'visibility' of the interference pattern produced by allowing the wanted line to illuminate an interferometer, he was able to deduce one term in the Fourier transform of the intensity–wavelength distribution in the line. By observing the 'visibility' for a large number of path differences in the interferometer he was therefore able to deduce the complete distribution in the spectral line.

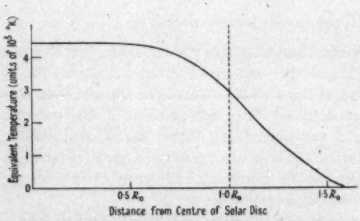


Figure 8. Distribution of 'brightness' across the solar disc at a wavelength of 60 cm.

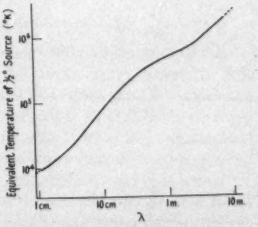


Figure 9. Variation of the emission from the undisturbed sun, as a function of wavelength. The intensity is expressed in terms of the equivalent temperature of a 'black' source subtending an angle of ½°.

Having obtained, in this way, the function $f(\theta)$, the intensity distribution of the equivalent 'strip' source, it is necessary to deduce the distribution across a source having circular symmetry (such as the sun) which would give rise to the observed 'strip' distribution.

Stanier made observations at a wavelength of 60 cm. using aerial spacings of up to 365 wavelengths; his results showed that the undisturbed sun appeared at this wavelength as a uniform disc having an equivalent temperature of $4{\cdot}5 \times 10^5$ deg. K. and of radius $0{\cdot}7R_0$, with a continuous decrease of 'brightness' at greater radii. The effective temperature at a radius of $1{\cdot}1R_0$ was half that near the centre of the disc, and was less than $0{\cdot}5 \times 10^5$ deg. K. at $1{\cdot}5R_0$ (Figure 8). The total intensity from the undisturbed sun corresponded to the emission from a source of angular diameter ½°, at a temperature of $5{\cdot}5 \times 10^5$ deg. K.

Before describing the observations of the sun which have been made at longer wavelengths, it may be worth summarizing the results obtained on wavelengths shorter than 1 m.

(i) In the absence of sunspots, the sun emits radio waves with an intensity which corresponds to a temperature of about 10^4 deg. K. at 1 cm., 8×10^4 deg. K. at 10 cm., 2×10^5 deg. K. at 25 cm. and $5{\cdot}5 \times 10^5$ deg. K. at 60 cm. (see Figure 9). In all cases the radiation appears to be emitted fairly uniformly over the visible solar disc, although weak radiation is detectable, especially at the longer wavelengths, from outside the visible disc.

(ii) In the presence of sunspots, an increase in intensity is observed which appears to be closely related to the sunspot area. The increased radiation is not of great intensity, and even at the longer wavelengths seldom exceeds 100% of that from the undisturbed sun. There is good evidence that the enhancement is due to emission from regions near the sunspots, and this radiation is frequently circularly-polarized.

(iii) Occasionally sudden, brief increases of intensity are observed, lasting for a few minutes only. The increases are small and seldom exceed 50% of the normal intensity. There is good evidence for their association with chromospheric flares.

(*b*) *Observations on Wavelengths greater than* 1 *m.*

The emission of radio waves of wavelength greater than 1 m. differs from that at shorter wavelengths by the relatively very much greater significance of sunspots. Thus, whilst the detection of the radiation from the undisturbed sun presents considerable difficulty, due partly to the masking effect of the galactic background and partly to the difficulty of constructing highly directional aerials, the presence of sunspots near the centre of the solar disc may give rise to intensities 10^4 times as great; under these conditions the power picked up by even a simple aerial may exceed the noise level of the receiver itself.

In the same way the sudden increases of still greater intensity ('outbursts') which are associated with chromospheric flares may produce an increase of intensity by a factor of 100 or more within a few seconds.

The very great intensities associated with sunspots make it a matter of some difficulty to determine the contribution of the undisturbed sun, and continuous recording over a period of many months is usually necessary before the true minimum intensity corresponding to the undisturbed sun can be determined.

The first determination of the emission from the undisturbed sun on the longer wavelengths was made by Pawsey (1946), who deduced that the intensity on a wavelength of 1·5 m. corresponded to the emission from a source having an angular diameter of $\frac{1}{2}^\circ$ at a temperature of a million degrees. Subsequent measurements by Ryle and Vonberg (1947) on wavelengths of 1·7 m. and 3·7 m. using the spaced-aerial interference technique described in §5 gave effective temperatures of 6×10^5 deg. K. and $1{\cdot}2 \times 10^6$ deg. K. respectively. Some recent experiments at Cambridge on a wavelength of 6·7 metres (Machin, unpublished) have shown that the undisturbed sun has an effective temperature of 2×10^6 deg. K. at this wavelength. Records obtained at Cambridge of the radiation from the undisturbed sun at wavelengths of 60 cm., 1·7 m. and 3·7 m. are shown in Figure 10.

There is little evidence available to indicate the distribution of intensity over the undisturbed solar disc at these longer wavelengths. Observations by Haikin

and Chikhachev (1948) on a wavelength of 1·5 m. during the total eclipse of May 1947 showed that at totality the intensity was still 40% of the uneclipsed value; during this experiment they observed, however, a relation between the intensity and the uneclipsed prominence activity, and it therefore seems that their results cannot be interpreted in terms of the distribution from the undisturbed corona.

During the passage of sunspots across the solar disc, very much more intense radiation is observed. The intensity may be as much as 100 times the normal intensity at wavelengths of 1–2 metres, or 10^4 times the normal intensity at the

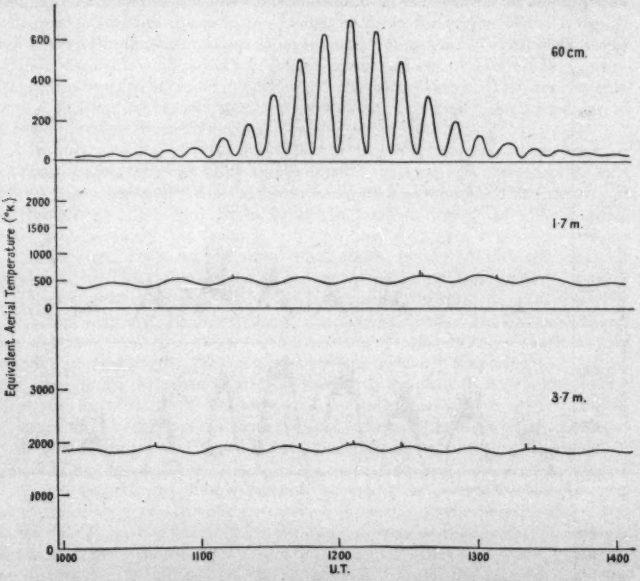


Figure 10. Records obtained with spaced-aerial systems, showing the radiation from the undisturbed sun on wavelengths of 60 cm., 1·7 m. and 3·7 m.

longer wavelengths. The radiation is usually a maximum when the sunspot is near central meridian passage, but the large irregular variations of intensity make an accurate determination of the 'emission polar diagram' of the source somewhat difficult. By studying a few of the largest sunspot groups, however, Appleton and Hey (1946b), and more recently Hey, Parsons and Phillips (1948b), concluded that the radiation on 4 metres falls off appreciably when the spot-group is more than 1–2 days from the centre. Similar conclusions were reached by Allen (1947) for radiation of 1·5 m. wavelength.

Recently Machin (unpublished) has carried out a correlogram analysis of the daily intensity of solar radiation recorded over a period of twelve months; this method of analysis does not attach undue weight to the larger sunspots, but

gives, effectively, the 'emission polar diagram' of an average sunspot. From records obtained on wavelengths of 1·7 m. and 3·7 m. he deduced that the intensity fell to half at angles of $\pm 15°$ and $\pm 7°$ respectively (corresponding to ± 1 day and $\pm \frac{1}{2}$ day from central meridian passage).

Many attempts have been made to correlate the intensity of the observed radiation with the size or type of the sunspots. Thus Pawsey, Payne-Scott and McCready (1946) concluded that there was a close relationship between the intensity of the 1·5 m. radiation and the total sunspot area, but more extended observations on a number of wavelengths have failed to show any simple relation.

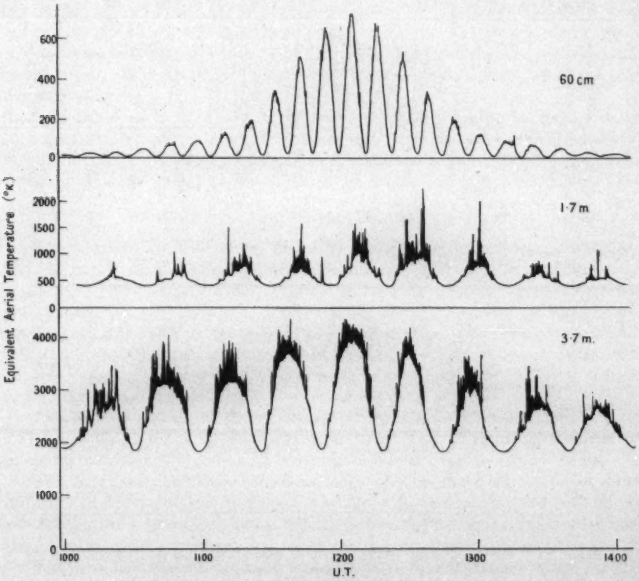


Figure 11. Records showing the solar radiation on wavelengths of 60 cm., 1·7 m. and 3·7 m. during the passage of a sunspot.

Apart from the very great intensity of the radiation, it is markedly different in character, as can be seen from some typical records shown in Figure 11. The rapid fluctuations of intensity which are characteristic of the enhanced radiation at wavelengths greater than 1m. do not show correlation between different wavelengths such as 1·7 m. and 3·7 m. (Ryle and Vonberg 1947).

Detailed examinations of the structure of some of these fluctuations have been made by Williams (1948) on a wavelength of 4 m. and Smith (unpublished) on a wavelength of 1·4 m.; their results have shown that many of the brief disturbances may double the intensity and have a duration of only about $\frac{1}{2}$ second.

The cause of the fluctuations is still unknown; in order to find out whether they could be due to some local refraction mechanism in the earth's atmosphere McCready, Pawsey and Payne-Scott (1947) made simultaneous observations at points 160 miles apart, and found close agreement in the records. It thus appears that the fluctuations are either due to variations in the emission from the source, or are caused by refraction effects in some distant region, such as the outer atmosphere of the sun above the source.

Although there is good evidence for assuming a relation between the occurrence of enhanced radiation and the presence of sunspots on the solar disc, it is important to determine whether the radiation is emitted from the region of the sunspot, or whether both phenomena are related to a common origin. Experiments using aerials of high resolving power were therefore carried out by McCready, Pawsey and Payne-Scott (1947) and by Ryle and Vonberg (1946). The experiments showed that the intense radiation was emitted from relatively small areas in the neighbourhood of the visible spot group.

The measurement of a maximum diameter of the sources of the enhanced radiation demonstrated that the emission from these regions corresponded to a still greater temperature than had previously been assumed. The intensity at wavelengths of 1·5–1·7 m. corresponded to temperatures of 10^9–10^{10} deg. K., whilst more recent measurements on 3·7 m. have suggested still greater values.

In addition to the normal rapid fluctuations associated with the emission from sunspots, there are occasionally very much larger increases of intensity which have been designated 'outbursts' by a number of workers. Disturbances of this type may involve an increase of intensity by a factor of 100 or more, and although the detailed structure of the variations of intensity may differ markedly on different wavelengths, the times of onset may be nearly simultaneous. These large and sudden disturbances are also frequently associated with chromospheric flares and disturbances to the terrestrial ionosphere which are connected with increased ultra-violet emission from the sun. A typical example of an 'outburst' which was observed at Cambridge on four different wavelengths is shown in Figure 12.

When disturbances of this type are observed on several wavelengths, the precise time sequence is a matter of some interest. Payne-Scott, Yabsley and Bolton (1947) have remarked that the time of onset of the shortest wavelength occurs first, as in the example shown in Figure 12; they therefore suggest that the disturbance is initiated by the elevation of prominence material from the photosphere through successive layers of the solar atmosphere. It is found, however, that such a time sequence, though frequent, is not invariant; observations at Cambridge on the same four frequencies as those in Figure 12 on 28th April 1949 showed a similar disturbance in which the order was exactly reversed. On other occasions an irregular order of onset has been observed, in which an intermediate wavelength was followed by both shorter and longer wavelengths.

The relation between 'outbursts' of radio-wave emission from the sun, the observation of chromospheric or prominence activity, and the occurrence of ionospheric storms is still somewhat obscure, as has been discussed by Allen (1947). There seems little doubt that the largest outbursts are nearly always associated with intense flare activity, but that for lesser disturbances the correlation is very much worse; many 'outbursts' do not appear to be related to visible disturbances, while many flares produce no observable increase in the radio

emission. It seems likely that the lack of correlation may be related to the angle over which radio waves can escape from a point in the solar corona. The propagation of radio waves in the solar atmosphere will be examined in detail in §9(*a*), but it can be noted here that radiation at the longer wavelengths is only likely to be able to escape from near a sunspot over a relatively small solid

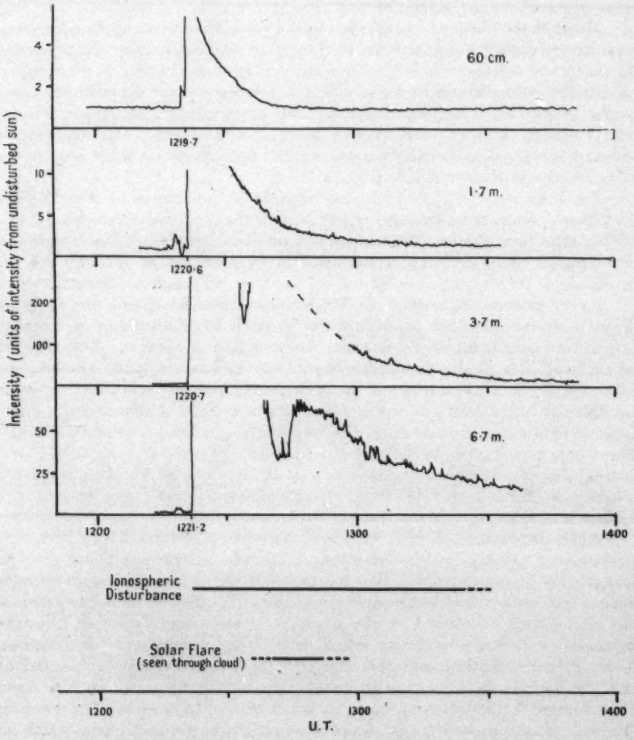


Figure 12. Sudden increase of intensity, or ' outburst ', observed on wavelengths of 60 cm., 1·7 m., 3·7 m. and 6·7 m. (1st February 1949).

angle; the observation that the normal enhanced radiation from sunspots shows a marked increase when the spot group is near central meridian is evidence for this theory. It might therefore be expected that sudden disturbances in regions near the limb would not necessarily produce intense radiation at the longer wavelengths; the observation by Lehany and Yabsley (1948) of ' outbursts ' on

the shorter wavelengths of 25 cm. and 50 cm. (where the conditions of escape are likely to be less rigorous) which were unaccompanied by disturbances on wavelengths of 1·5, 3 and 5 m. is additional evidence for such a mechanism. Hey, Parsons and Phillips (1948b), on the other hand, have found that for very intense 'outbursts' on a wavelength of 4 m. the correlation with chromospheric flares does not show a marked increase for flares occurring near the centre of the solar disc; it is possible that the conditions of escape of such intense disturbances are modified by large variations in the electron density in the solar atmosphere near the flare, and that escape of the radiation over a wider angle becomes possible.

Extensive observations have been made of the polarization of solar radiation in the band of wavelengths from 1 m. to 7 m. (Martyn 1946a, Appleton and Hey 1946a, Ryle and Vonberg 1946). It has been found that the radiation from the undisturbed sun is randomly polarized, but that the radiation associated with sunspots is usually circularly polarized; the sense of the polarization is different for different sunspots. Under conditions when the radiation is completely polarized, it is found that the rapidly fluctuating component associated with the enhanced emission from sunspots is also completely polarized, with the same sense as that of the steady component. At times when the radiation appears to be incompletely polarized, rapid fluctuations of both polarities are observed; it is possible that at these times radiation is received from two sources, each of which itself emits radiation which is fully polarized.

The state of polarization of the large 'outbursts' is not completely known; there is some evidence (Pawsey 1950) that the radiation is randomly polarized, although it has usually been found that the intense radiation shortly after such an onset is polarized. A conclusive experiment is difficult owing to the rapidly changing intensity at the beginning of such disturbances, and to the rare occurrence of such a short-lived phenomenon.

Although there is uncertainty about the polarization during these periods of extremely great intensity, there seems little doubt that the normal enhanced emission from sunspots is fundamentally circularly-polarized, although on occasions the presence of more than one source may reduce the degree of polarization. It is natural to associate a circularly-polarized wave with the occurrence of emission or absorption processes in the presence of a magnetic field, particularly since intense magnetic fields are known to occur in the neighbourhood of sunspots. It will be seen later in §9(*a*) that two possible regions of origin are to be expected for the emission from sunspots, and that these two regions will emit circularly-polarized radiations of opposite sense. A determination of the relation between the sense of the polarization and the direction of the magnetic field in the sunspot (as determined by visual observations) should therefore allow discrimination between these two possible regions. Owing to the difficulty of identifying the source of radio waves with a particular sunspot, the early observations (Ryle and Vonberg 1948) were not conclusive, although it was then suggested that the radiation corresponded to the 'ordinary' component of Appleton's magneto-ionic theory.

More recent experiments on wavelengths of 1·4 and 3·7 m. have been carried out (Stanier, unpublished) in which the position of the source of the radio waves was determined on each occasion by using a spaced-aerial system of large resolving power. Although it was possible to locate the source to an accuracy of 2 minutes

of arc by this means, the complex magnetic nature of most spot-groups still made it difficult to decide with certainty the direction of the magnetic field at a point high in the solar corona, where the source of the radiation is to be expected. Although the results can still not be regarded as conclusive, it is felt that on wavelengths of both 1·4 and 3·7 m. there is good reason for supposing that the radiation corresponds to the 'extraordinary' wave. It will be seen later (§ 9(*a*)) that this wave corresponds to a source at a greater altitude in the solar atmosphere than for the case of the 'ordinary' wave.

It is now possible to summarize the results of observations on the sun at wavelengths between 1m. and 20 m.:

(i) In the absence of sunspots the intensity of radio waves from the sun corresponds to the emission from a 'black' source of angular diameter $\frac{1}{2}^\circ$ at a temperature of about 6×10^5 deg. K. at 1·5 m., 10^6 deg. K. at 3·7 m. and 2×10^6 deg. K. at 6·7 m. wavelength (see Figure 9). No observations of the radiation from the undisturbed sun have yet been made at longer wavelengths.

The radiation appears to be randomly polarized.

(ii) In the presence of sunspots, greatly enhanced radiation is observed; at wavelengths of about 1·5 m. the intensity may be increased by a factor of 100. At longer wavelengths increases of as much as 10^4 have been observed. These increases are of quite a different order of magnitude from those observed on wavelengths shorter than 1 m. The enhanced radiation is characterized by rapid fluctuations of intensity, and is usually circularly polarized.

(iii) The enhanced radiation is emitted from comparatively small areas in the neighbourhood of the visible sunspots. The emission from these areas corresponds to temperatures greater than 10^9 deg. K. at 1·5 m. and 10^{10} deg. K. at 3·7 m. wavelength. No observations have yet been made to determine the diameter of the sources at longer wavelengths.

(iv) Occasionally still greater intensities are observed for brief periods, often accompanied by chromospheric flares or other signs of solar activity. Such disturbances are characterized by a very sudden onset. Little information is available on the polarization or dimensions of the source of this radiation.

§ 8. EXPERIMENTAL OBSERVATIONS OF GALACTIC RADIATION

(*a*) *The general Radiation from the Galaxy*

The remarkable observation that radio waves of measurable intensity were emitted by the galaxy was made by Jansky (1932) on a wavelength of 16 m. He was unable to detect any similar radiation from the sun, and this suggested that the galactic radiation could not be the integrated radiation from the stars, for which an intensity very much smaller than that from the sun would be expected. Later observations by Reber (1940, 1944) using a shorter wavelength (1·9 m.) and an aerial having a very much greater resolving power than that used by Jansky enabled the distribution of intensity to be plotted with considerable detail. His results confirmed that the radiation was emitted mainly from the plane of the galaxy, and showed the existance of maxima in the constellations of Sagittarius, Cygnus, Cassiopeia, Canis Major and Puppis.

More recent observations by Hey, Parsons and Phillips (1948a) (in which a special technique was used for correcting the observed intensities for errors caused by the finite width of the radiation pattern of the aerial system) enabled

a detailed distribution of the intensity to be obtained on a longer wavelength of 4·7 m. Their results again showed a general agreement with the structure of the galaxy, but attempts to obtain a closer agreement with the distribution of various types of star were unsuccessful. The distribution of intensity on a wavelength of 60 cm. was also determined by Reber (1948 b), and although the flux was too small to be measured except near the galactic plane, it is clear that the general distribution is also the same at this wavelength.

The most remarkable feature of these observations was the very great intensities which were observed. Fränz (1942), who has made measurements on a wavelength of 10 m., concluded that the intensity at this wavelength corresponded to the emission from a distributed source at a temperature exceeding 10^5 deg. K. The results obtained by Hey, Parsons and Phillips (1948 a) on a wavelength of 4·7 m. corresponded to a temperature of $1{\cdot}8 \times 10^4$ deg. K. in the direction of the centre of the galaxy. Thomas and Burgess (1947) have deduced that Reber's observations on 1·9 m. correspond to a maximum temperature of $1{\cdot}5 \times 10^3$ deg. K., while his observations on 60 cm. correspond to a maximum temperature of only about 30° K. No radiation from the galaxy on shorter wavelengths has yet been detected.

From these observations it can therefore be concluded that, except at the shortest wavelengths, the emission from the galaxy corresponds to a temperature which varies approximately as the square of the wavelength, and that on all wavelengths from 50 cm. to 10 m. the angular distribution of intensity is closely related to the general structure of the galaxy. More detailed work on the variation of effective temperature with wavelength has been carried out by Moxon (1946) for wavelengths between 1·5 m. and 7 m. He concluded that for directions of maximum intensity the effective temperature varied as $\lambda^{2{\cdot}7}$, whereas in directions near the galactic poles the relation was $\lambda^{2{\cdot}1}$. Whilst such a conclusion will be seen to be of great significance when considering the possible origin of the galactic radiation (§ 10(*d*)), it is important to bear in mind the difficulties and possible errors when making such a comparison with aerial systems of different directivity, particularly since it is now known that a considerable fraction of the galactic radiation is due to a few intense sources of small angular diameter. In another experiment using aerials of very small directivity, Herbstreit and Johler (1948) concluded that for the two wavelengths 12 m. and 2·7 m. the effective temperature varied as $\lambda^{2{\cdot}4}$.

(*b*) *The Discovery of discrete Sources in the Galaxy*

The observations described in the previous paragraph could have been accounted for by the emission from any type of source having a distribution throughout the galaxy, and as will be seen in § 10 (*a*), the emission by the interstellar gas at first appeared a more likely origin than the emission from stars. In 1946 however, Hey, Parsons and Phillips (1946) observed that the radiation on a wavelength of 5 m. from the direction of Cygnus sometimes showed rapid fluctuations of intensity within a period of a few minutes. They concluded that these fluctuations could not be caused by any refraction process in the earth's atmosphere, and suggested that this observation showed that at least part of the galactic radiation could not be due to a diffuse emitting region. The appearance of the fluctuations of intensity suggested strongly that the emission was analogous to that associated with sunspots.

Later observations using aerials of high resolving power were made by Bolton and Stanley (1948) on a wavelength of 3 m. and by Ryle and Smith (1948) on a wavelength of 3·7 m. The former workers used a 'Lloyd's mirror' interferometer involving an aerial mounted on a high cliff overlooking the sea, in the manner first used by McCready, Pawsey and Payne-Scott (1947) for locating the sources of intense sunspot radiation. By observing the variations of the recorded

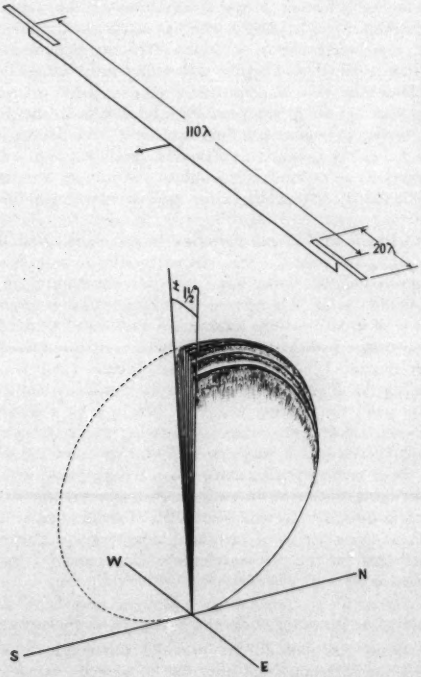


Figure 13. Diagram of aerial system and radiation pattern of the 'spaced-aerial' method of locating discrete sources in the galaxy.

intensity as a 'point' source rises above the horizon they were able to discriminate against the general background radiation, and by deducing the times of rising and setting they could determine both coordinates of its position; the East–West coordinate or Right Ascension was given by the mean of the times of rising and setting, whilst the North–South coordinate or declination was related to the difference between these two times.

Using this method they were able to locate the source in Cygnus with a greater accuracy than was obtainable by Hey, and they also showed that the angular diameter of the source was less than the resolving power of their apparatus (about 8 minutes of arc). Later observations (Bolton 1948) demonstrated the existence of a number of other discrete sources of smaller intensity.

Ryle and Smith (1948) used a spaced-aerial interferometer analogous to that used by Michelson for determining stellar diameters. In this method two aerials, each having a radiation pattern which is restricted in an East–West direction ($\pm 1\frac{1}{2}°$) but broad in a North–South direction ($\pm 45°$), are mounted on a line running East–West (see Figure 13). The resulting radiation pattern consists of a narrow 'fan' beam which is split up, by the interference between the two aerials, into a number of maxima separated by about $\frac{1}{2}°$. If the aerials are mounted so as to radiate vertically upwards at the latitude of Cambridge (52° N.), the 'fan' beam will extend from the north pole to within a few degrees of the equatorial plane; as the earth rotates, the radiation pattern will therefore be swept across nearly the whole of the northern hemisphere. In the absence of any discrete sources of radiation (having an angular diameter less than $\frac{1}{2}°$)

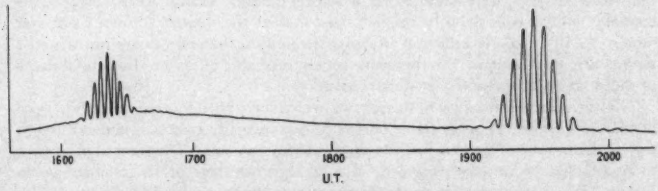


Figure 14. Section of record obtained with spaced-aerial interferometer, showing the intense sources in the constellations of Cygnus (1620) and Cassiopeia (1930).

the recorded intensity will remain constant, or will exhibit slow variations as the envelope of the radiation pattern is directed to different regions of the galaxy; if, however, the radiation pattern is swept across a discrete source of small angular diameter, the recorded intensity will show a periodic variation while the source is situated within $\pm 1\frac{1}{2}°$ of the meridian. By observing the time of the central maximum of this record, the Right Ascension or East–West coordinate of the source can be found. The North–South coordinate, or declination, may be determined by using another property of the spaced-aerial interferometer; the rate of change of path-difference from a remote source to the two aerials depends upon the declination, δ, of the source, and is proportional to $\cos\delta$. The periodicity of the recorded variation is therefore given by $t_0 \sec\delta$, where t_0 is the periodicity for a source situated in the plane of the equator. (t_0 is the time taken for the earth to rotate through an angle equal to the separation between the interference maxima and, for the aerial system shown in Figure 13, is about 2 minutes.)

This type of interferometer has advantages over the 'Lloyd's mirror' interferometer, since observations are made at meridian transit, and the angle of incidence of the radiation on the ionosphere and troposphere is a minimum; the possibility that errors might be introduced by refraction in the earth's atmosphere is therefore greatly reduced.

Using this system on a wavelength of 3·7 m., Ryle and Smith observed the source in Cygnus and also located another of still greater intensity in the constellation of Cassiopeia; they also located 28 sources of smaller intensity. A typical record showing the two intense sources is shown in Figure 14. None of the sources appeared to coincide with an outstanding visual star, and the two intense sources (for which an accuracy of position of 5–10 minutes of arc was possible) could not be associated with any star of greater than 8th magnitude.

(*c*) *The Distribution of the discrete Sources*

In view of the absence of correlation between the intense sources of radio waves and bright visual stars, it is important to investigate the possibility that the sources might be comparatively local bodies, situated within the solar system. Attempts have therefore been made to measure the distance of the two most intense sources by determining their parallax; accurate measurements of the positions of the sources in Cygnus and Cassiopeia were made, on a wavelength of 3·7 m., over two periods six months apart (Ryle and Smith, unpublished). The results of these observations showed that there was no change of the position of either source greater than the experimental accuracy; this result indicated that both sources were situated at distances greater than 2×10^{16} cm. This distance, whilst considerably smaller than that of the nearest known fixed star (about 3×10^{18} cm.), is sufficient to demonstrate that the sources are not situated within the solar system; it therefore seems probable that the discrete sources of radio waves are situated at stellar distances.

The angular distribution of the sources which have been located in the northern hemisphere shows no appreciable concentration near the galactic plane, although the total number of known sources is at present too small for great significance to be attached to this distribution. A map showing their positions is shown in Figure 15. A more or less uniform angular distribution would be expected to occur if the observed sources were at distances small compared with the dimensions of the galaxy (e.g. if they were at stellar distances).

It will be shown later (§ 10(*d*)) that on the supposition that the observed discrete sources of radio waves are situated at stellar distances, and that other sources are distributed in the same way throughout the galaxy, it is possible to account for the general background radiation from the galaxy in terms of the total radiation from all such sources; this theory does not require any mechanism based on the emission from the interstellar gas.

Since the relation between intensity and wavelength might be different for the emission from discrete sources and from the interstellar gas, additional experimental evidence for this hypothesis might be obtained by a comparison of the intensity–wavelength distribution of the discrete sources with that of the general galactic background. Observations of the two intense sources have therefore been made on a number of different wavelengths. Bolton and Stanley (1948) observed the source in Cygnus on wavelengths of 1·5, 3, 3·5 and 5 m., and concluded that over this range of wavelengths the intensity showed a shallow maximum at 3 m. Ryle and Smith (unpublished) observed the sources in Cygnus and Cassiopeia on wavelengths of 1·4, 3·7 and 6·7 m. and deduced the following figures for the intensity (watts/m^2/(c/s.)) $\times 10^{-22}$):

λ	1·4 m.	3·7 m.	6·7 m.
Cygnus	1·3	1·4	1·3
Cassiopeia	1·8	2·3	2·2

These results may be compared with those obtained by Moxon (1946) and Herbstreit and Johler (1948) for the general galactic background; their results suggested a variation of effective temperature which was between $\lambda^{2\cdot1}$ and $\lambda^{2\cdot7}$. These relations correspond to a variation of the flux per unit solid angle proportional to $\lambda^{0\cdot1}$ and $\lambda^{0\cdot7}$.

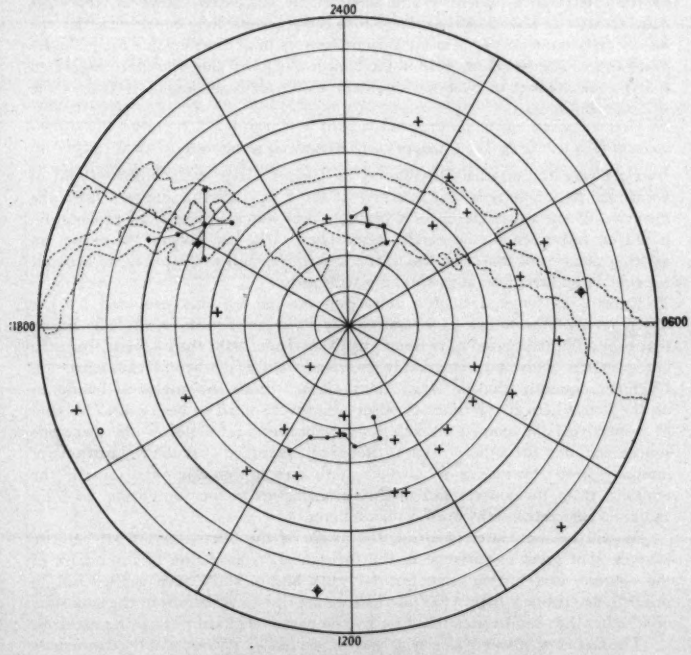


Figure 15. Map of northern hemisphere, showing the discrete sources located by Bolton and Stanley (○) and Ryle and Smith (+).

Until more accurate observations over a wider range of wavelengths and covering a greater number of sources have been made it is doubtful if the apparent difference between the two distributions of intensity with wavelength can be regarded as significant.

(*d*) *The Polarization of the Radiation from the discrete Sources*

It is natural to associate the intense emission from discrete sources in the galaxy with the enhanced radiation from sunspots. It is therefore important to examine the polarization of the radiation in order to determine whether it has the circular polarization characteristic of that from sunspots. Experiments by Ryle and Smith (1948) using the interference polarimeter described in §6

showed that the steady component of the radiation from the two intense sources, on a wavelength of 3·7 m., was unpolarized. Similar results were also obtained on wavelengths of 1·4 and 6·7 m. The measurement of the polarization of the variable component of the radiation from the source in Cygnus presented some difficulty, since the period of the fluctuations was comparable with the period of the interference pattern. The experiments suggested, however, that this component was also unpolarized. From these results they concluded that the steady emission from the discrete galactic sources might not involve the presence of an appreciable magnetic field in the emitting region; they therefore suggested that the mechanism of emission might be more nearly analogous to that of the undisturbed sun.

(*e*) *The Cause of the Fluctuations of Intensity*

Following the original observation by Hey, Parsons and Phillips (1946) of occasional fluctuations in the intensity of the 5 m. radiation emitted from the direction of the constellation of Cygnus, it was natural to seek an explanation based on refraction in the earth's atmosphere. It was difficult, however, on existing theories of the structure of the ionosphere and troposphere, to account for such large variations at so short a wavelength.

Further evidence against a refraction mechanism was provided by the observation of the source in Cassiopeia by Ryle and Smith (1948) who found that over an initial period of two months (May–June 1948) the intensity from the latter source remained remarkably constant, whilst that from the source in Cygnus frequently showed large fluctuations. Since the angles of incidence on the ionosphere at the times of observation were small in both cases (12° and 6° respectively), it seemed unlikely that atmospheric refraction could affect one source and not the other. It was therefore concluded that the fluctuations of intensity from the source in Cygnus were due to genuine variations of the emission from the source, and might be analogous to the 'outbursts' of solar radiation associated with chromospheric flares.

As will be seen later (§10(*b*)), the origin of the fluctuations from galactic sources is of great importance in the theoretical investigation of the nature of the sources, and further experimental work has recently been carried out in order to determine whether the fluctuations are due to variations in the emission, or whether they can be accounted for by comparatively local refraction processes.

The first experiment of this type was carried out by Pawsey and his co-workers (private communication): they observed the source in Cygnus simultaneously from 'cliff' receiving sites in Australia and New Zealand on a wavelength of 3 m. A comparison of the records obtained showed that on occasions when fluctuations were detected there was a marked absence of correlation between the variations received at the two sites. This result suggested that the fluctuations were caused by refraction in the earth's atmosphere.

As has already been pointed out, however, observations using the 'Lloyd's mirror' type of interferometer based on reflection at the surface of the sea involve observations of the source at a relatively small angle above the horizon; refraction effects in the ionosphere, and more particularly in the troposphere, are therefore likely to have a very much more serious effect than when the source is at a large elevation. Since additional refraction effects of this type might mask any genuine variations of the intensity, it appeared desirable to repeat

the experiment using a method which allowed observations to be made when the source was at a large angle of elevation.

Two series of experiments have therefore been carried out in this country, where observation of the source in Cygnus is possible with a very small angle of incidence (12°) and where additional evidence may be obtained by observing the source in Cassiopeia. The first experiment was carried out in cooperation with Dr. A. C. B. Lovell of the Jodrell Bank Experimental Station of the University of Manchester; in this experiment 3·7 m. receivers at Cambridge and Jodrell Bank (distance 210 km.) were used to make simultaneous observations of the source in Cygnus while it was near the meridian. In order to avoid the confusion caused by an interference system, a single aerial of considerable directivity was used, the intensity of the background radiation having been found previously from records using the interference system. Typical records obtained at Cambridge are seen in Figure 16, which shows the source in Cygnus during a

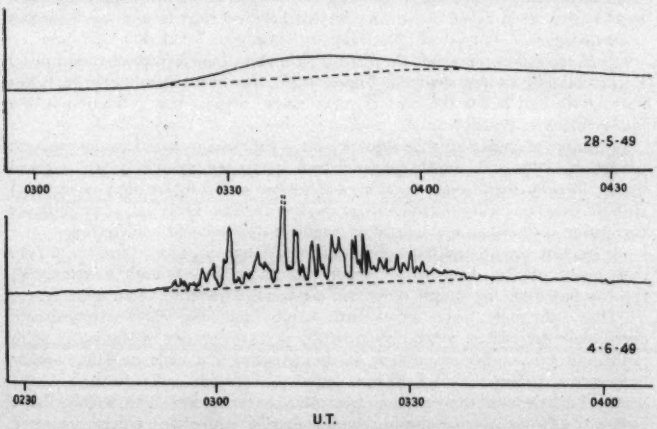


Figure 16. Records of the source in Cygnus, obtained with a single aerial: (*a*) 28th May 1949; (*b*) 4th June 1949.

'quiet' and a 'disturbed' period. The dotted line indicates the intensity of the background radiation.

In the second experiment, observations were made at a longer wavelength of 6·7 m., since it was expected that any refraction effects would be relatively greater than at the shorter wavelength. In these experiments simultaneous observations were made with one receiver at Cambridge and another which was moved to various sites at distances of up to 170 km. from Cambridge.

In both series of experiments it was found that although there was good correlation between the nights on which the records were disturbed, the greater part of the fluctuations observed at widely separated sites did not show detailed correlation. It was also found that whereas earlier experiments had suggested

that the source in Cygnus was markedly more active than that in Cassiopeia, during the period March–October 1949 both sources frequently showed large fluctuations of intensity. These results therefore suggested that the origin of the greater part of the fluctuations observed during this period was comparatively near to the earth.

In addition to the type of fluctuations shown in Figure 16(*b*), however, another type of variation was occasionally observed on 6·7 m. in which sudden large increases of the intensity occurred for a brief period of 10–20 seconds. The occurrence of this type of disturbance did not appear to be related to the first type, and simultaneous observations at two widely separated sites (Cambridge–Malvern) showed good agreement on the records. The occurrence of these disturbances was too infrequent, and their duration too brief, for their direction of origin to be definitely identified with that of either source, but it does not seem likely that they could have a terrestrial origin.

The present, incomplete, experimental evidence therefore suggests that two mechanisms are involved in the fluctuations of the intensity from discrete sources in the galaxy:

(i) A mechanism based on the disturbing effects of the terrestrial atmosphere: Records similar to that shown in Figure 16(*b*), in which the intensity fluctuates both above and below the normal mean value, suggest that refraction in the ionosphere is responsible.

In order to obtain further evidence for a mechanism based on ionospheric refraction, Ryle and Smith (unpublished) compared the periods of marked fluctuation (obtained from a series of observations on 3·7 m. lasting nine months) with the occasions when the ionosphere was known to be disturbed; no significant correlation with either ionospheric or magnetic storms could be detected.

It has also been found from observations at Cambridge on 3·7 m. and 6·7 m. that many of the disturbances occur simultaneously on both wavelengths; theories based on any simple refraction mechanism therefore seem unlikely.

Thus, although there seems little doubt that the earth's atmosphere occasionally introduces marked fluctuations in the intensity of the metre-wave radiation from the discrete sources, further experimental evidence will probably be necessary before the mechanism is found.

(ii) Variations of the emission from the source itself: The sudden large increases of intensity observed simultaneously with two widely separated receivers are suggestive of the 'outbursts' of solar radiation. As with the solar disturbances, however, their rare occurrence and brief duration make detailed study difficult, and further experimental evidence is necessary before their relation to solar 'outbursts' can be definitely established.

Both types of fluctuation are of considerable interest theoretically, the first because of its possible implications on the structure of the ionosphere, and the second because it enables deductions to be made concerning the physical dimensions of the source of the radiation.

§9. THEORIES OF SOLAR RADIATION

Experimental observations of the sun at radio wavelengths have been made by a large number of workers, and the results obtained have enabled a coherent, though incomplete, picture of the phenomena to be built up for wavelengths between 1 cm. and 10 m. The theoretical work, on the other hand, has not

reached a stage where there is general agreement; there are differences of particular importance in the theories dealing with the emission of the enhanced metre-wave radiation associated with sunspots. Some of these differences can be accounted for by the different conditions assumed in the solar corona (for which inadequate data are available from visual observations). Other differences are, however, fundamental, and their resolution presents one of the most important problems not only in Radio Astronomy but in Astrophysics generally. It is in this field that radio observations and deductions based on these observations should provide the data concerning stellar envelopes which are lacking in observations at visual wavelengths.

In this section an attempt will be made to put forward, in outline, the most important theories of the emission of radio waves from the sun; the extension of these theories to account for the emission from galactic sources will be considered in § 10.

The experimental observations of the emission from the sun were summarized in the final paragraphs of §§ 7 (*a*) and 7 (*b*).

Before the possible mechanisms for generating radio waves are discussed, it is important to consider their propagation in the solar atmosphere, with particular reference to the conditions necessary for their escape from the sun.

(*a*) *Application of the Magneto-ionic Theory to the Solar Atmosphere*

The magneto-ionic theory of Appleton (1932) has been applied to the propagation of radio waves in the solar atmosphere by Saha, Banerjea and Guha (1947), by Martyn (1948) and by Ryle (1948). The conclusions reached will now be discussed briefly.

In the absence of a magnetic field, the refractive index, μ, in a region containing N electrons per cm^3, for a wave of angular frequency p is given by

$$\mu^2 = 1 - \frac{4\pi N e^2}{\epsilon_0 m p^2},$$

where ϵ_0 represents the permittivity of free space and e and m are the charge and mass of the electron. Propagation of a wave of a given frequency is therefore only possible in regions where $N < \epsilon_0 m p^2 / 4\pi e^2$. Thus in the absence of a magnetic field, a wave of angular frequency 5×10^8 (corresponding to a wavelength of 3·7 m.) cannot be propagated in the solar corona at heights of less than $1{\cdot}5 \times 10^{10}$ cm. above the photosphere (see Figure 17 (*a*)).

In the presence of a magnetic field, the refractive index is modified, and for propagation along the direction of the field is given by

$$\mu^2 = 1 - \frac{4\pi N e^2}{\epsilon_0 m p(p + p_H)} \quad \text{(ordinary wave)}$$

$$\mu^2 = 1 - \frac{4\pi N e^2}{\epsilon_0 m p(p - p_H)} \quad \text{(extraordinary wave),}$$

where $p_H (= He/\epsilon_0 mc^2)$ is the natural frequency of gyration of an electron in the magnetic field. The two waves are circularly polarized in opposite senses (the wave having the same sense as the natural rotation of a free electron being known as the extraordinary wave).

It can be seen that the ordinary wave may be propagated whatever the value of p_H/p providing that $4\pi N e^2/\epsilon_0 m p^2 < 1$, and that for large values of p_H/p

propagation can occur even with larger values of N. The presence of an intense magnetic field therefore allows the propagation of one circularly-polarized component in a region which would normally be over-dense.

The extraordinary wave, on the other hand, cannot be propagated unless $p_H/p > 1$ or $< 1 - 4\pi Ne^2/\epsilon_0 mp^2$. For intermediate values of p_H/p this wave cannot be propagated.

The significance of a magnetic field can be demonstrated more clearly by applying these expressions to the propagation of 3·7 m. wavelength radiation in the solar atmosphere. In Figure 17 (*a*) the variation of μ^2 with height is shown in the absence of a magnetic field, from which it can be seen that propagation

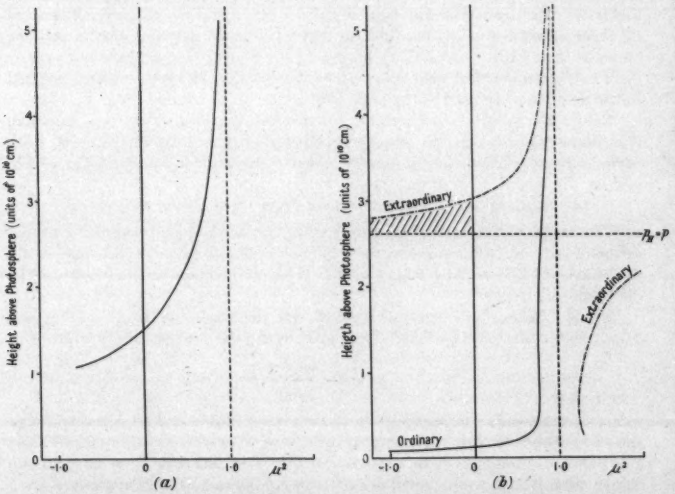


Figure 17. Curves showing the variation of μ^2 with height in the solar atmosphere for a wavelength of 3·7 m.: (*a*) no magnetic field; (*b*) above a 3,000 gauss sunspot.

can only occur at heights of more than $1{\cdot}5 \times 10^{10}$ cm. above the photosphere. Figure 17 (*b*) shows the effect of superimposing the magnetic field of a unipolar sunspot which is assumed to have the form of a dipole situated $0{\cdot}1R_0$ below the photosphere (R_0 = radius of photosphere) and producing a magnetic field of 3,000 gauss at the photosphere. (No account has been taken of any variation of the electron density over a sunspot in this example.)

Propagation of one component can now occur down to a height of 6×10^9 cm., but the extraordinary component cannot be propagated between $2{\cdot}6 \times 10^{10}$ cm. and 3×10^{10} cm. At a height of $2{\cdot}6 \times 10^{10}$ cm. the wave frequency and the natural frequency of free gyration of an electron are equal ($p_H = p$).

From this example it is clear that in the absence of an appreciable magnetic field the source of any 3·7 m. radiation which can reach the earth must be at a

height of more than $1{\cdot}5 \times 10^{10}$ cm. above the photosphere. The emission from a region above a sunspot could occur from a source situated above 6×10^{8} cm. (ordinary ray) or 3×10^{10} cm. (extraordinary ray).

In addition to determining the variation of refractive index for a particular wavelength in the solar atmosphere, it is also important to examine the absorption; it will be seen later that some of the theories for the emission of radio waves from the sun are based on 'equilibrium' mechanisms involving the emission from a heated electron gas. By the application of Kirchhoff's Law the emission may be determined if the absorption coefficient and temperature of the region are known. (Martyn (1948) has examined the validity of applying Kirchhoff's Law to cases where a magnetic field is present, and has concluded that in the case where the limiting polarizations are circular, as occurs in most problems of the solar atmosphere, Kirchhoff's Law may be applied.)

Ryle (1948) has shown that for the fully-ionized conditions of the solar corona, the absorption of metre-wave radiation is small except in three different regions. One of these regions occurs where the frequency of free gyration of an electron is equal to the wave frequency; this region of 'resonance absorption' will absorb the extraordinary component, and the absorption will be very great. Appreciable absorption will also occur in regions where the refractive index is small; each of the two circularly-polarized components will therefore undergo absorption at the heights at which the refractive index for that wave tends to zero.

The magnitude of the absorption coefficient for different wavelengths entering the solar atmosphere will be discussed later, in connection with the determination of the emission to be expected from a heated solar atmosphere. It may be noted here, however, that for a wave which approaches the surface of $\mu = 0$ normally the total absorption will be relatively great, whilst a wave approaching at a large angle of incidence will be reflected at a finite value of μ (given by Snell's Law), and the total absorption may be greatly reduced.

The preceding outline of the propagation of radio waves in the solar atmosphere has shown that any mechanism for the generation of radio waves must occur at a considerable height in the solar atmosphere. The examination by visual methods of the regions responsible for the generation of radio waves is therefore difficult, and the information which is available concerning the physical conditions in these regions is somewhat limited. For this reason it has not been possible to base the various theories of the emission of radio waves from the sun on generally accepted data; as will be seen later, it now seems probable that the radio observations themselves will enable more accurate deductions to be made concerning the structure of the solar atmosphere than are at present available from observations at visual wavelengths.

(*b*) *The Types of Mechanism suggested to account for the Emission of Radio Waves from the Sun*

In a recent communication (Ryle 1949 a) the theories of the emission of radio waves from the sun were divided into two groups:

(i) 'Equilibrium' mechanisms, in which the intensity of the emitted radiation tends to the value associated with the mean electron energy.

(ii) Non-thermal mechanisms, in which the coherent oscillation of a large number of electrons is capable of producing a very much greater intensity than would correspond to the mean energy of the electrons.

Deductions based on the first group require a knowledge of the mean electron energy or 'electron temperature' in the emitting region, together with a knowledge of the emissive power of the region. Theories based on this mechanism must also account for the maintenance of the very high electron temperatures which are required to produce the observed intensities.

In the second group a very high electron temperature in the region of the source is not required, but a detailed mechanism must be provided to account for the maintenance of the coherent oscillation of a large number of electrons.

(*c*) *Radiation from the Undisturbed Sun*

The experimental observations have shown that at wavelengths of the order of 3 cm. the undisturbed sun emits as if it had a temperature of about 2×10^4 deg. K. and at 50 cm. about 5×10^5 deg. K. At longer wavelengths (up to 7 m.) the emission corresponds to a gradually increasing temperature of up to 2×10^6 deg. K. (Figure 9). Experiments have suggested that at wavelengths of 60 cm. and shorter the emission is more or less uniform across the solar disc; no evidence is available for the distribution at longer wavelengths.

Theories of the emission based on equilibrium mechanisms have been proposed by Martyn (1948) and Ryle (1948); these theories made use of thermodynamic arguments to deduce the emission from a region whose absorption coefficient was known. Alternative methods of deducing the emission from a heated ionized gas have been used by Greenstein (1947), Unsöld (1947) and Waldmeier (1948). These authors have deduced the total emission by considering the radiation from an individual electron–electron or electron–proton encounter (theory of 'free–free' transitions). Martyn (1948) has shown that both these methods give the same result when the refractive index does not depart appreciably from unity, but that under conditions where the refractive index is small the theory of free–free transitions becomes complicated. Since many of the most interesting experimental observations have been obtained on wavelengths for which the refractive index in the solar corona is markedly different from unity, the theory has not been applied extensively; it has, however, been used to deduce the emission from the sun on the very short wavelengths, where the absorption is likely to occur in a region where the refractive index does not depart significantly from one.

The thermodynamic arguments used by Martyn (1948) and Ryle (1948) do not suffer from this difficulty, and are equally applicable to regions where the refractive index is very small; they are therefore particularly suitable for determining the emission at the longer wavelengths, for which the absorption is only appreciable where the refractive index is small. Both authors have deduced the total absorption in the region of the solar atmosphere where μ is small for an incoming wave, and have shown that for wavelengths between 1 and 10 m. the absorption is appreciable. They therefore concluded that the emission from the sun at each wavelength should be comparable with the emission from a black-body having the same temperature as that of the electrons in the region where μ is small.

On this basis, the measurement of the intensity on a number of different wavelengths should provide information concerning the electron temperature at different levels in the solar atmosphere (as was suggested by Ginsburg (1946)).

Denisse (1949 a) has raised objections to this treatment, concerning the relation between the absorption coefficient and the emissivity of a region in which the refractive index is very small. He maintains that for $\mu \simeq 0$, although the absorption coefficient becomes very large, the emissivity tends to zero. This conclusion appears to arise from a confusion between the true emissivity and the power transmitted across a boundary where the refractive index changes dis continuously from a very small value to one of unity. In the treatment of both Martyn and Ryle, the 'bulk' absorption and emission coefficients are used, and in the case of the solar corona, where the gradient of refractive index per wavelength is extremely small, the 'reflection loss' need not be considered.

Martyn (1948) has extended his analysis to the case of the absorption of a wave arriving at an oblique angle; in this way he was able to deduce the emission from different parts of the sun for wavelengths between 20 cm. and 30 m. In this analysis he assumed that the temperature of the corona was 10^6 deg. K. (from Edlén's (1942) identification of the spectra of highly ionized atoms in the corona) and that the temperature of the chromosphere was 30,000° K. (from measurements by Redman (1942) of the profiles of chromospheric spectral lines). He assumed a rather rapid variation of temperature at the boundary between the corona and the chromosphere.

His results indicated that at the longer wavelengths (greater than 10 m.) the total absorption even for normal incidence was not complete, and that partial reflection occurred. At these wavelengths the emission, even from the centre of the disc, was therefore somewhat less than that expected from a source at 10^6 deg. K. At wavelengths shorter than 1 m., on the other hand, a ray which is incident normally is not reflected until it has penetrated into the chromosphere; the intensity of the emission from the centre of the disc is therefore related to the temperature of the chromosphere. His theory therefore predicts that at wavelengths shorter than 1 m. the emission from the centre of the disc will be less than that from near the limb. Similar conclusions have been reached by Unsöld (1947) and Waldmeier (1948), who have based their calculations on the same assumptions regarding the temperature distribution in the solar atmosphere.

Observations at wavelengths of 60 cm., 10 cm. and 3 cm. have all suggested that the sun appears to radiate as a uniform disc, or as a disc showing appreciable 'limb-darkening'. It is probable that the discrepancies between the theoretical treatments of Martyn, Unsöld and Waldmeier and the experimental results are due to the assumption of an incorrect distribution of temperature in the solar atmosphere. It can be shown that the emissivity of the solar atmosphere for wavelengths of 10 cm. to 1 m. is rather critically dependent on the variation of electron temperature and density with height at the boundary of the corona and the chromosphere; it is in just this region that the information provided by visual observations is incomplete. It is probable that a more detailed understanding of the mechanisms involved in maintaining the electron temperatures in the corona and chromosphere would enable a more accurate picture to be built up for the dividing region. A possible mechanism for the maintenance of high electron temperatures in the solar atmosphere will now be described, and a more detailed examination of its effects in different regions will be given in § 9 (*e*). On this theory no sharp gradient of electron temperature between the corona and chromosphere would be expected, and the 'centre-darkening' predicted by Martyn, Unsöld and Waldmeier would not occur.

(d) The Production of high-energy Electrons in the Solar Atmosphere

It has been shown that, on the assumption that the electrons in the solar corona have kinetic energies corresponding to a temperature of about 10^6 deg. K., the emission of radio waves from the sun can be accounted for satisfactorily on the basis of the random motion of the electrons. Under certain conditions the intensity of the emitted radiation can approach the value associated with the emission from a black-body at a temperature of 10^6 deg. K.; for other conditions the intensity may be less, either because the total absorption of a wave is insufficient for equilibrium to be established, or because the radiation may be emitted by a lower level in the atmosphere, where the electron temperature is less. In addition to the evidence provided by radio observations, there is good evidence from visual observations that the temperature of the corona is about 10^6 deg. K. Waldmeier (1945) has summarized this evidence and has shown that four independent methods give comparable results for the coronal temperature:

(i) The relation between the pressure gradient in the corona and that in the photosphere corresponds to a coronal temperature about 1,000 times that of the photosphere.

(ii) Observations by Grotrian (1931, 1934) on the broadening of the photospheric lines scattered by electrons in the corona have indicated an electron temperature of 7×10^5 deg. K.

(iii) The observation of a large number of spectral lines whose ionization potentials exceed 100 volts (corresponding to the thermal energy at 10^6 deg. K.).

(iv) Observations of the width of the Fe XIV line have shown a Doppler broadening which would correspond to a temperature of 2×10^6 deg. K.

It is now important to examine possible mechanisms for maintaining a coronal temperature of about 10^6 deg. K. Ryle (1948) suggested that the electrons in the solar atmosphere were maintained at a high energy by the presence of an electric field. On the assumption that the observed rotational speeds shown by sunspots at different latitudes correspond to a true differential movement of the matter of the photosphere, Alfvén (1937) has shown that the presence of a general magnetic field would give rise to a potential difference between the poles and equator of the sun. Assuming a polar field strength of 50 gauss, this potential difference would be of the order of 5×10^8 volts.

In estimating the effects of such a field, it must be remembered that charged particles in the solar atmosphere can only move appreciably in a direction parallel to the magnetic field. If the magnetic and rotational axes of the sun were coincident, each line of magnetic force would cut the surface of the sun at points of equal potential, and particles in the solar atmosphere could not be accelerated to large energies by this mechanism. If, however, there is a small angle between the rotational and magnetic axes, a fraction of this potential could appear between two points on a line of magnetic force; for an angle of 7° (a figure which has been deduced from coronal observations at times of low sunspot activity) this potential is 10^6–10^7 volts, and in the absence of space-charge effects would give rise to a potential gradient of about 10^{-5} volts cm^{-1}, capable of accelerating charged particles in the chromosphere and corona. The mean electron energy which would be maintained by such a field depends on the distance travelled by an electron in the field before suffering an energy-losing collision; the electron temperature at different levels in the solar atmosphere will therefore depend

on the particle density and the types of collision at which energy can be lost. A more detailed analysis of the variation of electron temperature with height will be given in §9(*f*), but Ryle (1948) showed that in the undisturbed corona at the heights responsible for the emission of metre-wave radiation a mean electron energy of 10^2–10^4 electron volts should be maintained, corresponding to an electron temperature of 10^6–10^8 deg. K.

Although much detailed work is needed before the theory of the emission of radio waves from the undisturbed sun can be regarded as satisfactory, there now seems little doubt that the fundamental mechanism is an 'equilibrium' one, and that electrons of large energy exist in the solar corona. More precise agreement between the theoretical predictions and the observational results may possibly not be attained as long as coronal data derived from difficult experiments at visual wavelengths form the basis of the theoretical analysis. It is possible that with the development of more refined measuring techniques radio measurements may eventually provide more accurate information on the temperature distribution in the corona than is at present available from observations at visual wavelengths.

(*e*) *The Enhanced Radiation from Sunspots*

(i) *Theories based on 'equilibrium' mechanisms.*

Both Kiepenheuer (1946) and Denisse (1947 a, b) suggested that circularly-polarized radiation would be emitted by the free gyration of electrons in the magnetic field of a sunspot. The magnetic field in the chromosphere and corona has a suitable intensity for the generation of radio waves having frequencies similar to those observed, but there are serious difficulties in accounting for the radiation by such mechanisms. It was shown in Figure 17(*b*) that radiation generated in this way (having the polarization corresponding to the extraordinary wave) would have to travel a very considerable distance in a region where the refractive index was imaginary before it could escape. Apart from possible leakage of the exponentially attenuated 'evanescent' wave, no transfer of energy through such a region can occur.

Kiepenheuer (1946) neglected this attenuation completely, whilst Denisse (1947 a, b) concluded that if the gradient of magnetic field were large, it was not serious. Kiepenheuer, however, also deduced intensities which were very much greater than that corresponding to the electron temperature, without specifying any mechanism for maintaining a coherent motion of a number of electrons. His results were based on the addition of the radiation emitted by the electrons situated throughout a considerable depth, and did not include the effects of re-absorption; from simple thermodynamic arguments it can be seen that this neglect of the re-absorption is not valid, and that the intensity cannot exceed the value associated with the electron temperature.

In a later paper Ryle (1948) examined in some detail the possibility that appreciable energy could escape from the region of 'resonance absorption' where the wave frequency is equal to the frequency of free gyration of an electron in the magnetic field. He showed that whilst radiation generated by the gyration of electrons in the magnetic field of a sunspot could travel freely towards the centre of the sun, the attenuation of an outward-travelling wave was very great; he therefore concluded that such a mechanism could not account for the observed intensities. The observation that energy can be transferred towards the centre

of the sun, however, may be of importance, since the asymmetrical flow of momentum will give rise to an outward 'radiation pressure' on the electrons in a diverging magnetic field. This point will be discussed later in § 11 (*a*).

In addition to the region of 'resonance absorption', Ryle showed that the solar atmosphere above a sunspot should absorb an incoming wave at two other layers; each layer would absorb one of the circularly-polarized components. These two layers correspond to the regions where the refractive index for the two waves first becomes small; for example, for a wavelength of 3·7 m. and a sunspot field of 3,000 gauss the two layers are situated at heights of about 6×10^8 cm. (ordinary wave) and 3×10^{10} cm. (extraordinary wave)(see Figure 17 (*b*)). He showed that the total absorption in each of these regions depends on the collisional frequency and on the gradient of refractive index, and that in both regions the total absorption is likely to be considerable. Each of the regions should therefore emit a circularly-polarized wave having an intensity related to the temperature of the electrons in the region; on the assumption that the temperature of the solar atmosphere increases with height, it might therefore be expected that the predominant polarization observed at the earth would correspond to that of the extraordinary wave. It is, however, also necessary to consider the relative emissivities of the two regions, and the relative solid angles which the two sources subtend at the earth.

An accurate theoretical estimation of the intensity of the two components is complicated by the absence of sufficiently accurate data on the electron density and temperature above a sunspot, and on the difficulty of determining the effective collisional frequency for a fully ionized gas. Martyn (1948) has suggested that the absorption of a radio wave will depend on the type of collision proposed by Chapman and Cowling (1939), for which a cross section considerably greater than that for un-ionized hydrogen is obtained. A similar result was obtained independently by Ginsburg (1944) when considering the absorption of radio waves in the terrestrial ionosphere in the presence of a large number of ions.

If this collisional frequency is adopted, the total absorption for metre-wave radiation in both of the regions suggested by Ryle (1948) will be complete for radiation incident normally on the surface of $\mu = 0$. It is therefore concluded that each region should emit circularly-polarized radiation in a direction normal to the surface of $\mu = 0$, with an intensity corresponding to the electron temperature in the region. Machin (unpublished) has examined the variation of intensity of the radiation emitted in other directions, in an attempt to deduce the 'polar diagram' of a sunspot at different wavelengths. The calculation of the propagation of the waves through the overlying layers is difficult and involves a detailed knowledge of the distribution of electron density near a sunspot, but his preliminary results have given 'polar diagrams' in reasonable agreement with those observed experimentally.

The polarization of the radiation received at the earth will depend on the relative intensities of the two components. The intensity of each wave is related to the total effective solid angle subtended by the source and to the electron temperature in the region. The effective solid angle is obtained by deducing the emissivity of each small element of solid angle and deriving an effective solid angle for a 'black' source. Whilst this analysis is possible for a simple sunspot, and whilst it is then possible to make an estimate of the electron temperature in each of the two regions (in a manner which will be discussed in § 9 (*f*), it has

seldom been found possible to carry out the complete analysis for actual sunspot groups, owing to the complexity of their magnetic fields; it has not therefore been possible to obtain conclusive confirmation by comparing the results with the experimental observations, although both theory and experiment suggest that the extraordinary wave (corresponding to the source at greater altitude) is normally more intense.

It is now necessary to consider mechanisms for maintaining large electron temperatures in the solar atmosphere near sunspots. The experimental observations have demonstrated that small areas in the neighbourhood of sunspots emit metre-wave radiation with an intensity which corresponds to a temperature of 10^8–10^{10} deg. K. Various mechanisms for the maintenance of electron temperatures of this order of magnitude (corresponding to a mean electron energy of 10^4–10^6 electron volts) have been proposed, based on the acceleration of the electrons in an electric field.

According to Alfvén (1937), a potential difference of about 5×10^8 volts between the poles and the equator of the sun will be maintained by the interaction of the general magnetic field and the non-uniform rotation of the surface matter. In the absence of sunspots it was shown that a small fraction of this potential was available for the acceleration of charged particles in the solar atmosphere. If now the magnetic field of a sunspot is superimposed on the general magnetic field, a very much greater electric field becomes available for the acceleration of particles in the solar atmosphere. The combined magnetic field will have lines of magnetic force which cut the photosphere at points of widely different latitude, and Ryle (1948) has shown that a potential difference of about 10^8 volts* would initially be available for accelerating particles in the solar atmosphere near sunspots. The potential gradient may be modified by the presence of space charges, and also by the movement of photospheric matter in the sunspot, as will be considered in §9(*g*) in connection with the origin of 'outbursts' and chromospheric flares. Nevertheless he concluded that the mean electron energy at heights of more than 10^9 cm. in the corona would be increased to at least 10^6 electron volts (corresponding to an electron temperature of 10^{10} deg. K.).

Giovanelli (1947, 1948) and Denisse (1947 b, 1949 b) have suggested an alternative mechanism for producing an electric field in the solar atmosphere near a sunspot. From a consideration of the growth of the magnetic field of a sunspot they concluded that intense electric fields would be produced in its neighbourhood. For large and rapidly growing spots they deduced that the fields should be sufficient to accelerate electrons to energies of the order of 10^6–10^7 electron volts.

It has been shown that the experimental observations of the emission of intense radio waves from sunspots can be accounted for by 'equilibrium' mechanisms if electrons in the solar atmosphere near sunspots are assumed to have a large energy. Visual observations of the movement of the photospheric matter and of the magnetic fields in the sun have indicated that electric fields having an intensity sufficient to maintain the required electron energy are likely to occur.

* This figure was deduced on the basis of a general magnetic field of 50 gauss. Recent observations (Thiessen 1949) have suggested that the field may be only a few gauss. It will be seen in § 9 (*f*), however, that the mean free path of an electron in the solar corona is likely to be considerably greater than was previously assumed; adequate excitation should be maintained with a general field of only 1 gauss.

(ii) *Theories based on the coherent oscillation of electrons.*

Owing to the very great electrical temperatures which are required to account for the observed intensities of metre-wave radiation from sunspots, other authors have suggested non-thermal mechanisms; the emission of radiation having an intensity greater than that corresponding to the electron temperature requires some mechanism for maintaining the coherent oscillation of a large number of electrons.

Bruce (1946) has suggested that electrical discharges in the solar atmosphere could give rise to radio emission, in a manner analogous to that of the terrestrial lightning discharge. As has been pointed out by Martyn (1947), however, such a mechanism seems unlikely in a highly ionized region such as the solar atmosphere, where the electrical conductivity is always large.

Martyn (1947) and Shklovsky (1947) have both suggested that the coherent oscillation of electrons in the corona could be produced by a mechanism analogous to the plasma oscillations familiar in gas discharge tubes. Haeff (1948, 1949) has suggested that the interaction of streams of electrons in the solar corona could give rise to intense radio waves by the amplification methods which have been observed in electron tubes. Bailey (1948a, b, 1949) has shown that the presence of two intermingling streams is not necessary; space-charge wave amplification can occur in an ionized gas in the presence of a single drift velocity.

Ryle (1949a) has made a detailed analysis of these possible mechanisms, and has concluded that although the theories proposed may offer a satisfactory explanation for the phenomena observed in the small-scale conditions of discharge and electron tubes, there are great difficulties in applying them to the large-scale conditions of the solar corona. Thus, although the electron density at different heights in the solar corona is suitable for the generation of the wavelengths observed from the sun, there are serious difficulties in accounting both for the excitation of the oscillations and for the escape of radiation from the oscillating region.

On all the theories which have been proposed, the excitation of the oscillations depends upon the velocity modulation of a stream of electrons; the accounts which have so far been published have given a detailed analysis of the very simplified case of one or two streams of electrons, each composed of electrons having very nearly the same drift velocity (the case of interest when considering the application of the theory to the design of electron-wave amplifying tubes). Electrons in the solar atmosphere, however, will not remain within a narrow band of velocities even if they are accelerated by a uniform electric field, since they will make a large number of elastic collisions involving large angles of deflection in the region of acceleration; as will be shown in §9(*f*), it is probable that throughout the greater part of the solar atmosphere the distribution of velocities of the electrons may be closely represented by a large random distribution superimposed on a relatively much smaller drift velocity in the electric field. The application of the mechanisms proposed by Haeff and Bailey to such conditions, particularly in the presence of a magnetic field, appears to present very great difficulties.

Even if a mechanism existed for the maintenance of the oscillations, it is doubtful whether any energy so generated could escape to the earth; the region of generation of a plasma oscillation has, for the plasma frequency, a refractive index which is zero, and transfer of energy by electromagnetic waves cannot

occur. Whilst energy can be removed from the oscillating region by the movement of a bunched electron stream in the small-scale conditions of a discharge tube (as has been suggested by Armstrong (1947) and Neill (1949)), such a mechanism is improbable in the solar atmosphere where the electrons would have to travel a large number of wavelengths before reaching a region from which radiation could be propagated; in the absence of a very uniform drift velocity the electron stream would not remain bunched. It appears that similar arguments may be applied to the mechanisms proposed by Haeff and Bailey, although no detailed account has been given of the mechanism proposed for the escape of the energy; the emission of electromagnetic radiation relative to a coordinate system moving with the electron stream appears, however, to be closely related to the escape of radiation from a normal stationary plasma.

Whilst no conclusive arguments have yet been given concerning the origin of the enhanced solar radiation, there are very great problems associated with its generation by the coherent oscillation of a large number of electrons of comparatively small energy. These difficulties lie in the application of the theories of plasma oscillations and space-charge amplification to the conditions of the solar atmosphere, where the motion of the electrons will approximate to a random distribution of velocities, and where the gradients of electron density and magnetic field are very small. While the electron density in the solar corona has the right order of magnitude to account for radiation of the observed wavelengths, both the excitation of coherent oscillations and the escape of radiation from the oscillating region present great difficulties on the existing theories.

On the other hand, the generation of the radiation by 'equilibrium' mechanisms involving the random motion of electrons of large energy does not appear to present very great difficulties; on this theory different layers of the solar atmosphere are shown to be absorbing to radiation of different wavelengths, and regions where the absorption is considerable will also emit radiation whose intensity is related to the electron temperature. The maintenance of a mean electron energy of 10^6 electron volts (corresponding to an electron temperature of 10^{10} deg. K.) in the corona above a large sunspot does not present any greater problem than that associated with the maintenance of a mean electron energy of 100 electron volts (corresponding to an electron temperature of 10^6 deg. K.) in the undisturbed corona; the existence of the latter temperature is well established by visual observations. Both conditions would be expected to arise from the electric field produced by the observed non-uniform rotation of the sun, in conjunction with the general magnetic field.

(*f*) *Detailed Examination of the mean Electron Energy maintained by an Electric Field.*

In the 'equilibrium' mechanisms which have been proposed to account for the emission of radio waves from the sun, it has been suggested that the large mean electron energy in the corona is maintained by an electric field. Ryle (1948) has suggested that both the normal coronal temperature and the enhanced temperature near sunspots are due to an electric field which is maintained by the interaction of the general magnetic field of the sun and the non-uniform rotational velocity; Giovanelli (1947, 1948) and Denisse (1947b, 1949b) have suggested that the enhanced electron temperature near sunspots is due to an electric field maintained during the period of growth of the magnetic field of the

sunspot. It is therefore important to examine in detail the mechanism by which the application of an electric field to an ionized gas can affect the mean electron energy. It will be assumed that the gas is mainly hydrogen.

It is clear that the application of an electric field will accelerate both electrons and protons and, for a region of large dimensions, a condition of equilibrium will eventually be established in which the mean energy gained by a particle in moving along the gradient of potential is just balanced by the mean energy lost in collisions. The mean energy of the electrons (or the protons) which will be maintained by a given electric field therefore depends on the collisional processes involved.

For a gas which is not completely ionized (such as in the chromosphere) the most important mechanism by which an electron can lose energy will be inelastic collisions with un-ionized atoms, and for small degrees of ionization the electron energy cannot greatly exceed a value determined by the effective ionization potential of the gas. For a more completely ionized gas the inelastic collisions become rarer and a progressively increasing mean electron energy becomes possible. It is now necessary to consider the mechanisms by which the electrons can lose energy when the gas is fully ionized, although it should be noted that for a gas such as the solar atmosphere the presence of heavy atoms will still allow loss of energy by inelastic collisions as the mean energy of the electrons rises to values sufficient to produce multiple ionization. The presence of heavy atoms would therefore result in a more gradual transition from the 'incompletely ionized' conditions of the chromosphere to the 'fully ionized' conditions of the corona.

For a region which may be regarded as fully ionized (e.g. electrons and protons) an electron can only lose energy by elastic collisions with other electrons or with protons, and by radiation (Bremsstrahlung). It is clear that a large number of deflecting collisions will occur before the electron has lost appreciable energy, and the particle motion will therefore approximate to a random motion superimposed on a relatively much smaller 'drift' motion in a direction determined by the electric and magnetic fields. Under these conditions no net interchange of energy by electron–electron collisions need be considered, and the mean electron energy can only be limited by collisions with protons or by radiation losses.

The first of these mechanisms has been examined by Giovanelli (1949), who has considered the energy lost in elastic conditions with protons, assuming the collisional cross section deduced by Chapman and Cowling (1939). In the greater part of this analysis he has assumed that the mean energy of the protons corresponds to their thermal energy at the temperature of the photosphere (6,000°). As has already been pointed out, however, the protons will also be accelerated by the electric field, and since they cannot lose energy by collisions with other protons or with electrons, their mean energy is likely to be very considerable, and except under conditions of a very rapidly changing electric field it is likely to approach that of the electrons. The observations of Waldmeier (1945) have in fact shown that for the coronal line due to the ion Fe XIV, a Doppler broadening corresponding to a temperature of 2×10^6 deg. K. is found. It must therefore be concluded that the mean energy of the protons (and probably all positive ions) will be comparable with that of the electrons. Under these conditions an electron will, on the average, lose no energy by elastic collisions, and the mean free path

(for appreciable loss of energy) will be determined by the much smaller cross section associated with the radiation loss (Bremsstrahlung). Calculations by Machin (unpublished) have shown that under these conditions the mean electron energy in the solar corona would normally be limited only by the available potential. In stellar bodies of greater physical dimensions the effect may be of some importance; a given electric field might be capable of maintaining a considerably greater electron temperature than would be deduced from the theory of Giovanelli (1949).

The mean free path for a proton, for which the radiation loss is very small, will be still greater, and very high energies of protons (or heavier nuclei) are likely to occur in such stellar bodies.

(*g*) *Theories of the Rapid Fluctuations associated with Sunspots and the sudden large Increases associated with Chromospheric Flares*

No satisfactory explanation has yet been proposed to account for the rapid fluctuations of intensity associated with sunspot radiation. The absence of correlation between the fluctuations observed on different wavelengths can readily be explained by the difference in height of the regions responsible for the generation of the different wavelengths, but the short period of the fluctuations presents considerable difficulties.

Two main mechanisms appear possible:

(i) The refraction of the radiation in regions above that of generation —fluctuations might be introduced in a manner analogous to the twinkling of stars due to atmospheric irregularities. Simultaneous observations of the fluctuations at two widely spaced receiving sites (McCready, Pawsey and Payne-Scott 1947) showed identical records, and therefore suggested that no local refraction mechanism could be responsible; the dimensions of any refraction pattern produced by refraction in the outer layers of the solar corona, however, might be sufficiently large for the records at two points on the earth to be identical.

(ii) Variations in the emission from the source itself. These variations might be caused by changes in the electron temperature or density (and therefore emissivity) in the emitting region.

Observations by Williams (1948) on 4 metres have suggested that the individual increases of intensity appear to have a rapid rise and an exponential decay suggesting a momentary increase of the electron temperature; on the other hand some experiments by Smith (unpublished) on 1·4 metres indicated that the shape of the pulses (whose average width was $\frac{1}{2}$ sec.) was symmetrical.

In a later paper (Ryle 1949 a) it was suggested that even though electronic plasma oscillations are unlikely to exist in the solar corona, very low frequency 'sound-wave' oscillations of the ions (such as have been proposed by Tonks and Langmuir (1929)) might occur; the propagation of a disturbance of this type through the emitting region might give rise to variations of the electron temperature or density.

It is probable that no definite conclusion concerning the origin of the fluctuations will be possible until more detailed experimental observations are available.

The sudden very large increases of intensity associated with chromospheric flares, however, seem less difficult to account for. Bruce (1946), Martyn (1947), Giovanelli (1947) and Ryle (1948) have all suggested mechanisms involving sudden electrical discharges in the solar atmosphere. Martyn supposed that the turbulent motion of ionized prominence matter in the neighbourhood of a sunspot would produce large and intermittent electric fields. Giovanelli suggested that the variations in electric field were produced by the growth of the magnetic field of the sunspot, although Newton (1948) has since shown from an analysis of a large number of great flares that there is no close relationship between the occurrence of flares and the rate of growth of the magnetic field; on the contrary, he finds that the greatest flares occur during periods of maximum magnetic field, when the rate of change is a minimum. Ryle has suggested that the potential difference developed by the non-uniform rotation of the solar surface will give rise to electrical discharges along the lines of magnetic force, which will in turn produce considerable lateral motion of the photospheric matter in the sunspot. This motion will develop a 'back-E.M.F.' tending to decrease the potential gradient in the solar atmosphere (in the manner of an electric motor), and will thus reduce the average electron energy in the solar atmosphere below that expected for the full potential difference. If the motion of the photospheric matter is disturbed, the full potential may become available, and a greatly increased electron energy will occur temporarily at all points on the discharge path. The increase of visual and ultra-violet radiation caused by the enhanced excitation at low levels is regarded as being the origin of the chromospheric flares.

In these theories it is possible that the elevation of prominence material, which may occur as a result of the discharge, plays an important part in the increase of intensity (as was suggested by Payne-Scott, Yabsley and Bolton (1947) in explanation of the time-delay of 'outbursts' observed on different wavelengths). The elevation of prominence material may be important both from the point of view of producing a local increase in the electric field and in increasing the emissivity or the solid angle subtended by the source.

The absence of good correlation between 'outbursts' of radio waves and chromospheric flares can probably be accounted for by the 'polar diagram' of the radio waves; further analysis of the propagation of radio waves through the solar atmosphere above a sunspot may give an adequate reason for the observations, although the analysis is likely to be difficult.

§ 10. THEORIES OF GALACTIC RADIATION

(*a*) *Origin in the Interstellar Gas*

The earliest observations of radio waves from the galaxy, and the failure to detect similar radiation from the sun, showed that the galactic radiation could not be accounted for by the emission from a large number of stars like the sun. Calculations based on later observations of the sun show, in fact, that the observed galactic radiation has an intensity about 10^8 times that expected from such an origin. It must therefore be concluded that the radiation is either due to a type of stellar body whose radio emission is far greater than that of the sun, or to emission from the interstellar gas. Observations at visual wavelengths had suggested that the sun was likely to be a typical star, and earlier theories of galactic radiation were therefore based on the emission from the interstellar

gas. Thus Henyey and Keenan (1940), van de Hulst (1945), Greenstein, Henyey and Keenan (1946) and Shklovsky (1947) deduced the intensity to be expected at various wavelengths from the emission accompanying the acceleration of electrons in the field of other charged particles in interstellar matter ('free–free' transitions); they assumed that the interstellar gas was ionized hydrogen at a temperature of the order of 10,000° K. Whilst these calculations gave reasonable agreement with the observations at wavelengths shorter than 5 metres, it was shown by Townes (1947) that all the experimental observations could not be accounted for unless a temperature of at least 10^5 deg. K. were assumed. At the same time Woolley (1947) showed that, except in the neighbourhood of early-type stars, the temperature was only likely to be of the order of 1,000° K.

In a later analysis Unsöld (1948) showed that since the variation of intensity with direction is approximately the same at all wavelengths shorter than about 7·5 m., the galaxy is 'optically thin' at these wavelengths; the total absorption for a wave traversing a distance equal to the dimensions of the galaxy is insufficient for equilibrium between the mean particle energy and the radiation to be established. The observed intensities could only be accounted for by assuming an electron temperature of at least 10^5 deg. K. and an electron density of 1·5 cm^{-3}.

It is thus clear that there are very great difficulties in accounting for the observations on the basis of emission from the interstellar gas; the observation of discrete sources in the galaxy not only proves the existence of another mechanism, but also suggests that this second mechanism may account for the whole of the observed radiation. As will be seen later, there is as yet no experimental evidence which contradicts the hypothesis that the whole of the radiation originates in discrete sources which are assumed to be distributed throughout the galaxy; the theories which have been put forward to account for the emission from discrete sources will therefore next be examined.

(*b*) *The Nature of the Sources*

The early observations of the source in Cygnus demonstrated the existence of an intense source of small angular diameter whose emission showed rapid variations of intensity; it was therefore natural to conclude that the radiation was emitted by a star in which mechanisms analogous to sunspots and chromospheric flares occurred (Hey, Parsons and Phillips 1948a). Later observations by Ryle and Smith (1948) showed that, unlike the emission associated with sunspots, the radiation was not circularly polarized. They therefore suggested that the emission corresponded more closely to that from the undisturbed sun. The constant mean intensity from day to day (disregarding the short-period fluctuations) which has been observed from the two most intense sources over a period of eighteen months also suggests that emission from a sunspot mechanism is unlikely.

Observations by Bolton and Stanley (1948) and Ryle and Smith (1948) have shown that the angular diameter of the source in Cygnus is less than about 6 minutes of arc; deductions based on measurements of the parallax and of the period of the fluctuations of the intensity have suggested that the angular diameter is very much smaller (Ryle 1949b). Recent comparisons of the fluctuations observed with two receivers spaced 150–200 km. apart have suggested that the greater part of the fluctuations are due to refraction effects comparatively near to the earth. The measurements have shown, however, that one component of

the fluctuations, usually in the form of isolated 'bursts' of duration 10–20 seconds, agrees closely at two such spaced receivers. The observations therefore suggest that whilst some local mechanism, which is not yet fully understood, causes variations of the received signal, there is an additional fluctuation which appears to be inherent in the emission from the source. The existence of short-lived 'bursts' of radiation from a distant source is of great significance, since it suggests that the dimensions of the source of this component do not greatly exceed the distance travelled by an electromagnetic wave in 20 seconds; the observations therefore suggest that the sources are of stellar dimensions.

Observations of the parallax of the sources in Cygnus and Cassiopeia by Ryle and Smith (unpublished) have shown that both sources are situated at a distance greater than 2×10^{16} cm. An analysis by Smith (unpublished), based on the relative intensities of the 25 most intense sources, which will be discussed later, suggests that the distance of the nearest source is likely to be comparable with the distance of the nearest visual star (about 3×10^{18} cm.).

Using the maximum diameter of 20 light-seconds, derived above, and the minimum observed distance of 2×10^{16} cm., it can be shown that the observed intensity of the 3·7 m. radiation corresponds to the emission from a source at a temperature of at least 10^{12} deg. K.; if the source is assumed to be at the distance of the nearest known star the emission corresponds to a temperature of at least 10^{14} deg. K.

(*c*) *Mechanisms for the Emission of Intense Radiation from the Sources*

It is now necessary to examine the various mechanisms which have been proposed to account for the emission of intense radio waves from the sun, in order to see whether they can account for the much greater emission from galactic sources. Since none of the galactic sources has been identified with a known visual star, the only evidence concerning the nature of the sources is that derived from the radio observations. Ryle (1949b) has attempted this analysis, and has shown that the conditions of electron density and magnetic field in the region of the source are similar to those in the solar corona; he therefore concluded that the arguments put forward against the theory that the intense solar emission is due to the coherent oscillation of a number of electrons are equally applicable to the case of emission from the galactic sources. He therefore suggested that the emission is due to the random motion of electrons having a mean energy of at least 10^{10} electron volts.

By applying an analysis due to Alfvén (1937), Ryle then considered the electric field which would be maintained in the atmosphere of a star having a magnetic field and rotational velocity considerably greater than those of the sun. He showed that for stars having a polar field strength of 5,000 gauss and a peripheral velocity of 2×10^{7} cm. sec^{-1} (as deduced by Babcock (1947)), the electric field should be sufficient to produce an emission of radio waves having an intensity about 10^{4} times that from the sun. Owing to the non-linear relation between the electron temperature and the applied electric field which is likely to occur in stellar bodies whose physical dimensions exceed those of the sun (see § 9(*f*)), it is possible that the emitted radio waves would have an intensity appreciably greater than this factor.

On the assumption that other stars, which have not yet been examined visually, have magnetic fields and rotational velocities somewhat higher than those observed by Babcock, he concluded that a mean electron energy of 10^{10} electron

volts (corresponding to an electron temperature of 10^{14} deg. K.) could be maintained in the regions responsible for the emission of 3·7 m. radiation. The existence of an electric field of this magnitude would also accelerate protons and other positively-charged particles to a very great energy, and some of these particles would be able to escape from the stellar atmosphere; the possibility that the discrete sources of radio waves are also responsible for the origin of cosmic rays (Ryle 1949 b) will be discussed later.

The observation that none of the 30 sources observed in the northern hemisphere appears to coincide with an outstanding visual star may imply the existence of a new type of stellar body in which a high-temperature corona is associated with a small visual temperature or opacity; the mechanism proposed for maintaining a high coronal temperature is not connected with any thermal sources of energy in the denser regions of the star.

An alternative mechanism for the emission from discrete sources has recently been suggested by Bolton, Stanley and Slee (1949), who pointed out that the position of one of the sources coincides, within the limits of experimental accuracy, with that of the Crab Nebula. This body is believed to consist of an expanding shell of gas resulting from an early supernova, and subtends at the present time an angle of about 5 minutes of arc. On the assumption that the electron temperature was about 2×10^6, and that radiation from the whole body could reach the earth, they concluded that the observed intensity at 3 m. wavelength could be accounted for. This theory does not, however, account for the simultaneous observation of very brief increases of intensity at widely separated receivers, and as will be seen later, a theory based on the presence of unusual stellar bodies may lead to difficulties in accounting for the distribution of the observed sources.

(*d*) *The Origin of the Total Galactic Radiation*

The number of discrete sources which has so far been located has been limited by the sensitivity and resolving power of the apparatus. There is no doubt that improvements in the apparatus would lead to the discovery of more sources, and it is now important to consider whether the total radiation from the galaxy is due to a large number of discrete sources distributed throughout the galaxy, only the nearest of which can at present be resolved, or whether there is also a contribution from the interstellar gas.

Measurements of the intensity emitted by the two major sources at wavelengths of 1·4, 3·7 and 6·7 m. (Ryle and Smith, unpublished) have suggested a wavelength–intensity law similar to that observed from the general galactic background; it is therefore possible that the background radiation could be accounted for on the basis of the emission from similar discrete sources, if the total number of such sources distributed throughout the galaxy were sufficiently large. Observations in the northern hemisphere have shown the presence of 30 sources whose angular distribution does not appear to be related to the general structure of the galaxy. This result suggests that the observed sources are at distances considerably smaller than the dimensions of the galaxy; the total number of such sources in the galaxy might therefore be sufficiently large to account for the intensity of the background radiation. In addition, it is found that the ratio of the intensity emitted by the most intense source compared with that emitted by the whole sky is of the same order as the ratio of the intensity of the brightest visual star to the total visual emission from stars over the whole sky.

Whilst this evidence cannot be regarded as conclusive, it suggests that further examination of the available data is justified, particularly in view of the difficulties of accounting for appreciable radiation by mechanisms based on the emission from the interstellar gas.

On the assumption that the whole of the observed background radiation is due to discrete sources, it is possible to make deductions concerning their spatial distribution by carrying out an analysis of the intensity of each of the known sources relative to the total integrated intensity due to the whole of the northern hemisphere. Starting with the present limited experimental data, Smith (unpublished) has attempted an analysis based on a simple model in which the galaxy is supposed to be spherical (of radius R), and to contain a uniformly random distribution of sources of equal absolute magnitude. If the average density of the sources is ρ, then the ratio of the power received from the nth most intense source, p_n, to that received from all sources, p_t (i.e. the general galactic background), is given by

$$p_n/p_t = R^{-1}(36\pi\rho)^{-1/3}n^{-2/3}.$$

If now $p_n^{-3/2}$ is plotted against n, a straight line should be obtained, whose slope enables $R\rho^{1/3}/p_t$ to be found. Since p_t may be measured, ρ, the average density of the sources, may be deduced if R is known. Before describing the results of this analysis it is necessary to examine the simplifying assumptions which have been made:

(i) It is clear that in any practical model the density of sources ρ will be a function of both distance and direction. Since only very near discrete sources are likely to be detected, however, it can be assumed that ρ is constant over the region containing the observed sources. The intensity of the background radiation, on the other hand, will depend markedly on the distribution of sources in the galaxy.

If the density of the sources at a point (r, θ, ϕ) is given by $\rho_0 f(r, \theta, \phi)$, where ρ_0 is the density of sources in the neighbourhood of the solar system, then the intensity of the radiation received from a small element of volume at this point is

$$A.\frac{\rho_0 f(r,\theta,\phi) r^2 \cos\phi \, d\theta \, d\phi \, dr}{r^2},$$

where A represents the emission per unit solid angle from a single source.

The total intensity received from the whole galaxy is then given by

$$A\rho_0 \iiint f(r,\theta,\phi)\cos\phi \, d\theta \, d\phi \, dr.$$

This intensity can be equated to the total intensity received from an equivalent spherical galaxy having a uniform density of sources ρ_0 and a radius R:

$$4\pi A\rho_0 \int_0^R \frac{r^2\,dr}{r^2} = 4\pi A\rho_0 R.$$

If, therefore, $f(r, \theta, \phi)$ is known, it is possible to derive R from the relation

$$4\pi R = \iiint f(r,\theta,\phi)\cos\phi \, d\theta \, d\phi \, dr.$$

The experiments have shown that the angular distribution of intensity of the background radiation is closely related to the structure of the galaxy as determined by visual observations. It is therefore reasonable to assume that the

distribution of the density of the sources of radio waves is similar to that of the visible stars, and by deducing $f(r, \theta, \phi)$ for visible stars a value of R may be determined.

(ii) In the simple analysis it was assumed that all the sources were of equal absolute magnitude. The existence of a wide range of magnitudes would modify the deduced separation of the sources. Until a more complete theory of the emission from the discrete sources is available it is impossible to predict what range of absolute magnitudes is likely to occur; it can be shown, however, that even if there were a large range of absolute magnitudes, such as occurs for visible stars, the average spacing of the sources would not be altered by a factor of more than about 10. It therefore seems possible to deduce the order of magnitude of the spacing without any knowledge of the range of absolute magnitudes of the sources.

In applying the analysis to the experimental observations it is also necessary to take account of the limited solid angle covered by the aerial system; in these observations the coverage was restricted to the northern hemisphere. The result of plotting $p_n^{-2/3}$ against n for the 25 most intense sources observed in the northern hemisphere on a wavelength of 3·7 m. is shown in Figure 18. From

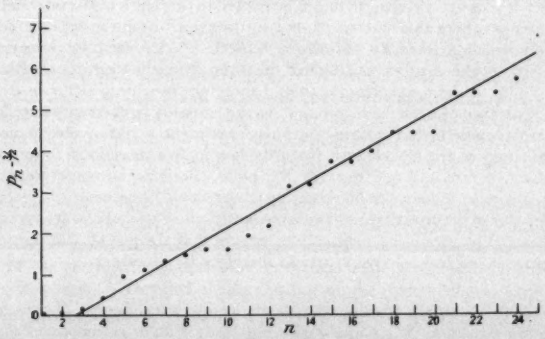


Figure 18. Plot of $p_n^{-2/3}$ against n for the 25 most intense sources observed in the northern hemisphere. (p_n is the power received from the nth most intense source, and is expressed in units of 10^{-24} watts/m^2/(c/s.).)

the slope of this graph the measured value of p_1 for the northern hemisphere and the approximate value of R deduced by the method described (1,000 parsec.) it was found that the average density of the sources was of the order of three per cubic parsec. This density corresponds to an average separation of the sources of about 0·7 parsec. (2×10^{18} cm.), a figure which is comparable with the average separation of the visible stars.

Whilst it is clear that this analysis is based on very limited experimental data, it can be concluded that the present observations are consistent with the theory that the whole of the background radiation of the galaxy is due to the emission from discrete sources similar to those observed in the constellations of Cygnus

and Cassiopeia, and that these sources are distributed throughout the galaxy with a density which is comparable with that of the visible stars.

In view of the very great difficulty of accounting for more than about 1% of the observed intensity of galactic radiation by theories relating to the emission from the interstellar gas, the theory of distributed discrete sources has much to recommend it. It is to be hoped that further experimental work will enable a definite conclusion to be reached concerning the two mechanisms. One possible way of obtaining definite proof for the existence of an interstellar mechanism involves the accurate determination of the angular distribution of intensity at a number of different wavelengths. If the radiation is due to an interstellar mechanism, the very great effective temperatures observed at the longer wavelengths (10 m.) in the direction of the galactic centre can only be accounted for by assuming that the 'optical depth' at these wavelengths is appreciable; the total absorption occurring in the depth of the galaxy is then too great for the intensity to be directly proportional to the depth of interstellar gas. At a shorter wavelength, on the other hand, the effective temperature is much less, a result which shows that the 'optical depth' is small. The angular variation of the intensity at a short wavelength would therefore be directly related to the depth of the interstellar gas in different directions; the angular variation of the intensity at a longer wavelength might be related to the depth of the interstellar gas in directions where this was small, but in directions near the galactic centre a relatively smaller intensity would be expected. An accurate experimental comparison of the angular variations of the intensity at the longest possible wavelength and at a short wavelength, should therefore provide conclusive evidence for the existence of a mechanism based on the emission from the interstellar gas. An alternative method of carrying out this experiment would involve an accurate determination of the wavelength–intensity law in two directions, one towards the galactic centre and one towards the pole. Both experiments are likely to be difficult, because considerable resolving power would be necessary to determine the true value in the direction of the galactic centre, and observations at a wavelength of at least 10 m. would be necessary to ensure that the galaxy was appreciably absorbing; a large aerial system would therefore be necessary.

§11. FUTURE DEVELOPMENTS IN RADIO ASTRONOMY

(*a*) *The Study of the Solar Atmosphere*

It has been shown that the extension of astronomical observations to radio wavelengths is the logical method of exploring regions of the universe where the density of matter is small. The existing experimental methods have already shown the possibility of investigating the structure of the solar corona, and although the theoretical interpretation of the results is still incomplete, there is no doubt that the new information obtained will be of fundamental importance in understanding the mechanisms involved.

One of the most important results of some of the existing theoretical work is related to the restricted band of wavelengths which will be emitted by any particular region of the solar atmosphere; observations on a particular wavelength should therefore allow deductions to be made concerning the conditions in a layer whose density is calculable. By making simultaneous observations over a range of wavelengths it should therefore ultimately be possible to build up a

complete picture of the temperature distribution at all heights in the solar atmosphere, with a detail which has no parallel in observations at visual wavelengths. This facility is due to the reflecting properties of a fully ionized atmosphere at radio wavelengths.

The extension of these observations to the disturbed conditions above sunspots should give important information on the mechanisms involved in the formation of prominences and chromospheric flares, and in the emission of corpuscular streams.

An examination of the emission of radiation by the gyration of the electrons in a magnetic field shows that whilst energy can be propagated into a region of increasing magnetic field strength, propagation in the reverse direction cannot occur. In the solar atmosphere above a sunspot the negative gradient of magnetic field strength will result in emission from each region downwards, towards the photosphere. The upward 'radiation pressure' on the emitting electrons may be of the same order of magnitude as the gravitational attraction on a hydrogen atom (Ryle 1948). A more detailed analysis of this mechanism, and a study of the electric fields likely to occur in the neighbourhood of sunspots, may throw light on problems related to the cooling of the photosphere in sunspots and the support of prominence matter.

(*b*) *The Study of Stellar Envelopes*

At the present time radio observations of individual stellar bodies are limited by the poor resolving power and sensitivity of the apparatus. The observations have shown that certain bodies, none of which has yet been definitely associated with a known visual star, but which appear to be of stellar dimensions, are capable of emitting radio waves whose intensity corresponds to a temperature exceeding 10^{14} deg. K.

Theories of the emission of intense radio waves from the solar corona were greatly assisted by the evidence available from visual observations; until radio waves can be observed from an identified visual star, theories concerning the nature of the sources must be based on the evidence available from the radio observations alone. Detailed measurements of the two major sources have, however, made it possible to deduce a maximum density for the emitting region —a figure comparable with that of the solar corona. By applying arguments used in theories of the emission from the solar corona, it was concluded that the intensity of the emission from the galactic sources implied the existence of a mean electron energy of at least 10^{10} electron volts.

The existence of mechanisms capable of maintaining a mean electron energy of this order of magnitude in the atmosphere of a stellar body would be of considerable importance, not only because of its relationship to the origin of cosmic rays, but also because of its connection with the structure of the atmospheres of visible stars. Whilst the lack of correlation between the intense sources and bright visual stars suggests that certain bodies emit an abnormal intensity of radio waves, deductions based on the total radiation from the galaxy suggest that they are not unusual bodies, but merely bodies in which the ratio of radio emission to visual emission is greater than that of the existing known stars. The existence of a high temperature in the atmosphere of a star is unlikely to be due to the same mechanisms as govern the photospheric temperature, as

can be appreciated from solar observations. The possibility of deducing details of the significant mechanism from observations of the intense radio sources may resolve some of the difficulties associated with certain visual stars.

(*c*) *The Origin of Cosmic Rays*

Mechanisms for producing electric fields in the atmospheres of rotating magnetized bodies have been suggested by several authors (Swann 1933, Alfvén 1937, Babcock 1948) in connection with the generation of cosmic rays. These mechanisms have been discussed in relation to the emission of intense radio waves from stellar atmospheres (Ryle 1949b). It was there shown that stars having characteristics which have been measured by visual observations (Babcock 1947) would be likely to emit radio waves having an intensity at least 10^4 times that from the sun. On the assumption that other stars have somewhat greater magnetic fields or rotational velocities, it was shown that it was possible to account for both the emission of radio waves having the observed intensity and the acceleration of charged particles to cosmic-ray energies. It was therefore suggested that the discrete sources of radio waves in the galaxy might be responsible for the primary cosmic rays. The possibility of such an origin has also been suggested by Denisse (1949b) and Unsöld (1949).

The observation of heavy particles in the primary radiation (Freier, Lofgren, Ney and Oppenheimer 1948) has provided good evidence that cosmic rays are due to the acceleration of charged particles in an electric field, whilst recent observations by Auger, Daudin, Denisse and Daudin (1949) have suggested strongly that the source of the primary radiation does, in fact, lie within the galaxy. In these experiments primary particles of very high energy were observed, and it was found that a significant sidereal variation of their intensity could be detected; this variation showed good agreement with the angular distribution of intensity of the general background of radio waves from the galaxy.

(*d*) *The Structure of the Galaxy*

In addition to the investigation of the nature of the sources themselves, it is likely that even with existing experimental techniques, important information could be obtained on the general structure of the galaxy. The absorption of visual radiation by dust particles obscures certain regions of the galaxy and makes the accurate delineation of distant regions inaccurate; dust particles introduce negligible absorption to radio waves, and even with existing resolving powers, important information might be derived from an accurate plot of the angular distribution of intensity. Attempts at such a survey have already been made (Hey, Parsons and Phillips 1948a, Northcott and Williamson 1948), but it is unlikely that they represent a very accurate distribution of the background radiation, owing to the neglect of the radiation from the major discrete sources.

In conclusion, it may be said that the existing investigation of the radio waves emitted by the sun and galaxy should only be regarded as a preliminary survey; the limits of sensitivity and resolving power of existing apparatus are not fundamental, and the development of more ambitious equipment should produce a marked increase in the available evidence. At the same time, the complications of the various mechanisms associated with regions of low density, which at the present time have prevented the universal acceptance of any one theory, should become clarified by further theoretical work.

ACKNOWLEDGMENTS

The preparation of this paper represents the combined work of the Radio Astronomy team at the Cavendish Laboratory. In addition to the several references to the unpublished work of Messrs. Machin, Stanier and Smith, the author would like to acknowledge the assistance provided by Messrs. Hewish and Elsmore. The work at the Cavendish Laboratory was made possible by a grant from the Department of Scientific and Industrial Research.

Figures 2, 3 and 5 appeared in a paper published in the *Proceedings of the Royal Society*, and permission to reproduce them here is gratefully acknowledged.

REFERENCES

Alfvén, H., 1937, *Z. Phys.*, **105**, 633.
Allen, C. W., 1947, *Mon. Not. R. Astr. Soc.*, **107**, 386.
Appleton, E. V., 1932, *J. Instn. Elect. Engrs.*, **71**, 642; 1945, *Nature, Lond.*, **156**, 534.
Appleton, E. V., and Hey, J. S., 1946 a, *Nature, Lond.*, **158**, 339; 1946 b, *Phil. Mag.*, **37**, 73.
Armstrong, E. B., 1947, *Nature, Lond.*, **160**, 713.
Auger, P., Daudin, J., Denisse, J., and Daudin, A., 1949, *C.R. Acad. Sci., Paris*, **228**, 1116.
Babcock, H. W., 1947, *Astrophys. J.*, **105**, 105; 1948, *Phys. Rev.*, **74**, 489.
Bailey, V. A., 1948 a, *Nature, Lond.*, **161**, 599; 1948 b, *Austral. J. Sci. Res.* A, **1**, 351; 1949, *Phys. Rev.*, **75**, 1104.
Bolton, J. G., 1948, *Nature, Lond.*, **162**, 141.
Bolton, J. G., and Stanley, G. J., 1948, *Nature, Lond.*, **161**, 312.
Bolton, J. G., Stanley, G. J., and Slee, O. B., 1949, *Nature, Lond.*, **164**, 101.
Bruce, C. E. R., 1946, *Observatory*, **66**, 263.
Burgess, R. E., 1941, *Proc. Phys. Soc.*, **53**, 293; 1946, *Ibid.*, **58**, 313.
Chapman, S., and Cowling, T. G., 1939, *Mathematical Theory of Non-Uniform Gases* (Cambridge : University Press), p. 177.
Christiansen, W. N., Yabsley, D. E., and Mills, B. H., 1949, *Nature, Lond.*, **164**, 569.
Covington, A. E., 1947, *Nature, Lond.*, **159**, 405; 1948, *Proc. Inst. Radio Engrs.*, **36**, 454; 1949, *Ibid.*, **37**, 407.
Denisse, J., 1947 a, *Ann. d'Astrophys.*, **10**, 1; 1947 b, *C.R. Acad. Sci., Paris*, **225**, 1358; 1949 a, *Ibid.*, **228**, 751; 1949 b, *Ibid.*, **228**, 467.
Dicke, R. H., 1946, *Rev. Sci. Instrum.*, **17**, 268.
Dicke, R. H., and Beringer, R., 1946, *Astrophys. J.*, **103**, 375.
Durand, E., Oberly, J. J., and Tousey, R., 1949, *Astrophys. J.*, **109**, 1.
Edlén, B, 1942, *Ark. mat. Astr. Fys.*, **28B**, H.1.
Fränz, K., 1942, *Hochfreq. tech. u. Elektroakust.*, **59**, 143.
Freier, P., Lofgren, E. J., Ney, E. P., and Oppenheimer, F., 1948, *Phys. Rev.*, **74**, 213.
Ginsburg, V., 1944, *J. Phys. U.S.S.R.*, **8**, 253; 1946, *C.R. Acad. Sci. U.R.S.S.*, **52**, 487.
Giovanelli, R. G., 1947, *Mon. Not. R. Astr. Soc.*, **107**, 338; 1948, *Nature, Lond.*, **161**, 133; 1949, *Phil. Mag.*, **40**, 206.
Greenstein, J. L., 1947, *Observatory*, **67**, 22.
Greenstein, J. L., Henyey, L. G., and Keenan, P. C., 1946, *Nature, Lond.*, **157**, 805.
Grotrian, W., 1931, *Z. Astrophys.*, **2**, 106; 1934, *Z. Astrophys.*, **8**, 124.
Haeff, A. V., 1948, *Phys. Rev.*, **74**, 1532; 1949, *Ibid.*, **75**, 1546.
Haikin, S. E., and Chikhachev, B. M., 1948, *Bull. Acad. Sci. U.S.S.R.*, **12**, 38.
Henyey, L. G., and Keenan, P. C., 1940, *Astrophys. J.*, **91**, 625.
Herbstreit, J. W., and Johler, J. R., 1948, *Nature, Lond.*, **161**, 515.
Hey, J. S., 1946, *Nature, Lond.*, **157**, 47; 1949, *Mon. Not. R. Astr. Soc.*, **109**, 179.
Hey, J. S., Parsons, S. J., and Phillips, J. W., 1946, *Nature, Lond.*, **158**, 234; 1948 a, *Proc. Roy. Soc.* A, **192**, 425; 1948 b, *Mon. Not. R. Astr. Soc.*, **108**, 354.
Houtgast, J., and Laffineur, M., 1948, *C.R. Acad. Sci., Paris*, **227**, 717.
Jansky, K. G., 1932, *Proc. Inst. Radio Engrs.*, **20**, 1920; 1933, *Ibid.*, **21**, 1387.
Kiepenheuer, K. O., 1946, *Nature, Lond.*, **158**, 340.

LAFFINEUR, M., MICHARD, R., STEINBERG, J. L., and ZISLER, S., 1949, *C.R. Acad. Sci., Paris*, **228**, 1636.
LEHANY, F. J., and YABSLEY, D. E., 1948, *Nature, Lond.*, **161**, 645.
LEHMANN, G., 1946, *Ann. Telécommun.*, **1**, 92.
MARTYN, D. F., 1946 a, *Nature, Lond.*, **158**, 308; 1946 b, *Ibid.*, **158**, 632; 1947, *Ibid.*, **159**, 26; 1948, *Proc. Roy. Soc.* A, **193**, 44.
MCCREADY, L. L., PAWSEY, J. L., and PAYNE-SCOTT, R., 1947, *Proc. Roy. Soc.* A, **190**, 357.
MOXON, L. A., 1946, *Nature, Lond.*, **158**, 758.
NEILL, T. R., 1949, *Nature, Lond.*, **163**, 59.
NEWTON, H. W., 1948, *Observatory*, **68**, 29.
NORTHCOTT, R., and WILLIAMSON, R. E., 1948, *J. R. Astr. Soc. Can.*, **42**, 269.
NYQUIST, H., 1928, *Phys. Rev.*, **32**, 110.
PAWSEY, J. L., 1946, *Nature, Lond.*, **158**, 633; 1950, *J. Instn. Elect. Engrs.*, Pt. III.
PAWSEY, J. L., PAYNE-SCOTT, R., and MCCREADY, L. L., 1946, *Nature, Lond.*, **157**, 158.
PAYNE-SCOTT, R., YABSLEY, D. E., and BOLTON, J. G., 1947, *Nature, Lond.*, **160**, 256.
REBER, G., 1940, *Astrophys. J.*, **91**, 621; 1944, *Ibid.*, **100**, 279; 1946, *Nature, Lond.*, **158**, 945; 1948 a, *Proc. Inst. Radio Engrs.*, **36**, 88; 1948 b, *Ibid.*, **36**, 1215.
REDMAN, R. O., 1942, *Mon. Not. R. Astr. Soc.*, **102**, 140.
RYLE, M., 1948, *Proc. Roy. Soc.* A, **195**, 82; 1949 a, *Proc. Phys. Soc.* A, **62**, 483; 1949 b, *Ibid.*, **62**, 491.
RYLE, M., and SMITH, F. G., 1948, *Nature, Lond.*, **162**, 462.
RYLE, M., and VONBERG, D. D., 1946, *Nature, Lond.*, **158**, 339; 1947, *Ibid.*, **160**, 157; 1948, *Proc. Roy. Soc.* A, **193**, 98.
SAHA, M. N., 1946, *Nature, Lond.*, **158**, 549.
SAHA, M. N., BANERJEA, B. K., and GUHA, U. C., 1947, *Indian J. Phys.*, **21**, 199.
SANDER, K. F., 1947, *Nature, Lond.*, **159**, 506.
SCHULKIN, M., HADDOCK, F. T., DECKER, K. M., MAYER, C. H., and HAGEN, J. P., 1948, *Phys. Rev.*, **74**, 840.
SHKLOVSKY, J. S., 1947, *Nature, Lond.*, **159**, 752.
SMITH, R. A., 1949, *Aerials for Metre and Decimetre Wavelengths* (Cambridge: University Press).
SOUTHWORTH, G. C., 1945, *J. Franklin Inst.*, **239**, 285.
STANIER, H. M., 1950, *Nature, Lond.*, **165**, 354.
SWANN, W. F. G., 1933, *Phys. Rev.*, **43**, 217.
THIESSEN, G., 1949, *Z. Astrophys.*, **26**, 16.
THOMAS, H. A., and BURGESS, R. E., 1947, *Survey of Existing Information and Data on Radio Noise over the Frequency Range 1 to 30 Mc/s.* (London: His Majesty's Stationery Office).
TONKS, L., and LANGMUIR, I., 1929, *Phys. Rev.*, **33**, 195.
TOWNES, C. H., 1947, *Astrophys. J.*, **105**, 235.
UNSÖLD, A., 1947, *Naturwissenschaften.*, **34**, 194; 1949, *Nature, Lond.*, **163**, 489.
VAN DE HULST, H. C., 1945, *Ned. Tijd. v. Nat.*, **11**, 210.
WALDMEIER, M., 1945, *Mitt. d. Aarg. Nat. Ges.*, **22**, 185; 1948, *Experientia*, **4**, 64.
WILLIAMS, S. E., 1948, *Nature, Lond.*, **162**, 108.
WOOLLEY, R. V. D. R., 1947, *Mon. Not. R. Astr. Soc.*, **107**, 308.

PLANETARY AND SATELLITE ATMOSPHERES

By GERARD P. KUIPER
Yerkes Observatory, University of Chicago

CONTENTS

	PAGE
§ 1. Introduction	247
§ 2. Criteria on the presence of a planetary atmosphere	247
§ 3. Spectroscopic observations of planetary atmospheres	250
§ 4. Visual and photographic observations of planetary atmospheres	253
§ 5. Observational results and interpretation	257
Mercury	257
Venus	258
Mars	261
Jupiter	266
Saturn	268
Uranus and Neptune	271
Pluto	272
General remarks	272
Appendix	272
References	273

§ 1. INTRODUCTION

PROGRESS in our factual knowledge of planetary atmospheres is slow and difficult. Spectroscopic investigations, with a few exceptions, are limited to the spectral range covered by photographic plates and to the region $\lambda > 3000$ A., transmitted by our atmosphere. Often very low dispersions must be used. Direct observations of details on the planetary discs are hampered by their smallness and resultant lack of resolution. In favourable cases the diameter of the disc is about 100–300 times the resolving power, the latter being set by the degree of turbulence in the terrestrial atmosphere through which the observations are made. Progress on the empirical side is therefore largely dependent on increased sensitivity of spectrographic receivers used on the largest telescopes, and on the development of observing sites having an unusually steady and transparent atmosphere. Even so we must expect our factual knowledge to remain very incomplete. Recourse must be had to indirect methods, based on physics, chemistry and meteorology, to complete the picture presented by the observations.

The observational data at hand are reviewed below. For greater detail the reader is referred to a monograph on the subject published recently (Kuiper 1949 a) as well as to other works mentioned in the course of the discussion.

§ 2. CRITERIA ON THE PRESENCE OF A PLANETARY ATMOSPHERE

(*a*) The spectrum of a planet may reveal absorption or emission bands, not caused by the passage of the light through our own atmosphere or by the sun itself;

(*b*) clouds, seasonal polar caps, twilight arcs or atmospheric refraction may be observed;

(*c*) indirect criteria may suggest the presence of an atmosphere: polarization phenomena not readily explained by the solid surface; a gradual darkening toward the planetary limb; a strong ultra-violet reflection, unexpectedly high in comparison with that in the visible spectrum, suggesting Rayleigh scattering by molecules;

(*d*) the velocity of escape on the surface of the planet (computed from its mass and radius), together with its computed equilibrium temperature based on the distance to the Sun, may show either that all normal gases are necessarily absent (because they would rapidly escape from the planet if present), or that the presence of gases with a molecular weight exceeding a given value is dynamically possible. This theoretical criterion is of great help in eliminating from the study hopeless cases, and in directing attention to the promising ones.

In view of its importance we shall first develop criterion (*d*) in greater detail. Subsequently we shall arrange the 9 planets and 30 satellites in order of their ability to retain an atmosphere. It will appear that all 9 planets and 7 of their satellites are objects deserving further study; they are listed in Table 1. These 16 objects are thereupon subjected to the criteria (*a*) to (*c*) above.

The velocity of escape is given by the condition that the kinetic energy per unit mass equals the potential energy per unit mass on the surface of the planet,

$$\tfrac{1}{2}v_e^2 = GM/R, \qquad \ldots\ldots(1)$$

in which M and R are the mass and radius of the planet, and G the constant of gravitation. The root-mean-square velocity of the molecules of molecular weight μ, called v_m, is given by

$$(v_m)^2 = \overline{v^2} = 3RT/\mu. \qquad \ldots\ldots(2)$$

Jeans has shown that if $v_m < 0{\cdot}2v_e$, the atmosphere is practically stable against evaporation for astronomical periods ($>10^9$ years); if $v_m > 0{\cdot}25v_e$, any atmosphere would dissipate so rapidly that, in view of the accepted age of the solar system, 3 to 4×10^9 years, no atmosphere would now be left.

The application of this criterion is somewhat complicated by the fact that the uppermost strata of a planetary atmosphere are expected to be hotter than corresponds to the black-body equilibrium temperature based on the intensity of the received solar radiation. Since the actual evaporation of the atmosphere takes place from the level at which the mean free path of the molecules in a vertical direction approaches infinity, it is the kinetic temperature of these layers that determines the rate of escape. For the Earth the temperature of this layer may be as high as 1,500–2,000° K., as compared to about 300° K. for the surface. Such high upper-air temperatures are due to the fact that the ultra-violet radiation of the sun is absorbed by a minute fraction of the total atmospheric mass. The temperature ratio between the uppermost atmosphere and the ground may be somewhat larger for the earth than for other planets because of the presence of free oxygen (Spitzer 1949) (O_3 and O are powerful absorbers for a wide range of ultra-violet radiations); nevertheless, most atmospheric constituents absorb powerfully below 1800 A., so that similar, if less extreme, conditions are expected to be found on other planets. Therefore the application of criterion (*d*) will over-estimate the stability of atmospheres if surface temperatures are used. In the later discussion some planets or satellites will be included which will in fact have no atmosphere.

Table 1 is a revision of a similar table published before (Kuiper 1949 b). A new mass determination of Mercury (Rabe 1949) removes this case from the very uncertain class; the correction to the previous value was 23%. The diameter of Pluto has now been measured with the 200-inch Hale telescope and was found to be so small that the published mass determinations of Pluto cannot be regarded as real; it will take at least another 20 years of observations before the mass can be determined directly. The diameter of Neptune is now known with sufficient precision (Kuiper 1949 d), at 0·90 of the earlier value. The diameter of Triton has appeared too small for direct measurement with the 82-inch telescope, but should be accessible to the Hale telescope.

Table 1. Planets and Principal Satellites: Physical Data*

	Mass	Radius	Mean density	Albedo	v_e (km/sec.)	r	$v_e r^{1/4}$ (km/sec.)
Mercury	0·0543	0·39	0·92	0·063	4·2	0·387	3·3
Venus	0·814	0·967	0·90	0·73	10·4	0·723	9·6
Earth	1·000	1·00	1·00	0·29	11·3	1·000	11·3
Mars	0·108	0·532	0·70	0·15	5·1	1·524	5·7
Jupiter	318·35	10·97	0·241	0·44	61·0	5·203	92
Saturn	95·3	9·03	0·13	0·42	36·7	9·539	64
Uranus	14·58	4·00	0·23	0·45	21·6	19·19	45
Neptune	17·26	3·50	0·40	0·49	25·1	30·07	59
Pluto	< 0·10	0·46	< 1·0	0·17	< 5·3	39·52	< 13
Moon	0·0123	0·273	0·607	0·07	2·4	1·000	2·4
J I	0·0121	0·264	0·66	0·53	2·4	5·203	3·7
J II	0·0079	0·236	0·60	0·55	2·1	5·203	3·1
J III	0·0260	0·391	0·44	0·35	2·9	5·203	4·4
J IV	0·0160	0·360	0·34	0·14	2·4	5·203	3·6
Titan	0·0235	0·373	0·45	0·28	2·8	9·539	5·0
Triton	0·022	0·35 ?	0·51 ?	0·28 ?	2·8 ?	30·07	6·6 ?

* Sources: Russell, Dugan and Stewart, *Astronomy*, Vol. I, 1945, Kuiper (1944), Ross (1928), Danjon (1948) and present paper (Pluto).

The masses, radii and mean densities given in Table 1 are expressed in terms of the earth. The albedo is the fraction of visible solar radiation reflected by the planet. The velocity of escape is computed from equation (1). It would immediately give the relative stability of planetary atmospheres if they had common T and μ values (equation (2)). Since the distances to the Sun differ through a range of 100 to 1, the energies received per cm² will have a range of 10^4, and the equilibrium temperatures a range of 10. The mean molecular velocities, for a given μ, will therefore have a range of $10^{1/2}$. It is clear that we can take out this effect on the relative stability by multiplying v_e of Table 1 by $r^{1/4}$, in which r is the mean distance to the Sun. This has been done in the last column of Table 1. Actually, the factor will be accurate only if the planets are entirely comparable in all other respects (composition of atmosphere, period of rotation, obliquity of the equator, etc.), but the approximation is a useful one, particularly since little is known *a priori* about μ. Therefore the quantities in the last column of Table 1 define, to a good first approximation, the order of ability for a planet or satellite to retain an atmosphere.

The order established by Table 1 is as follows: the Jovian planets, Earth, Venus, Pluto, Triton, Mars, Titan, J III, J I, J IV, Mercury, J II, and the Moon. (J I–IV are the Galileian satellites of Jupiter.) If we had included the remaining 23 satellites we would have found Rhea next, but with only half the value of $v_e r^{1/4}$ found for the Moon. On the Moon only the heaviest atmospheric gases would be stable dynamically ($\mu > 32$ if $T = 300°$ K., and $\mu > 41$ if $T = 107°$ C. $= 380°$ K.); Rhea would require $\mu > 150$. Therefore Table 1 includes all the bodies in the solar system on which atmospheres might conceivably be found.

§ 3. SPECTROSCOPIC OBSERVATIONS OF PLANETARY ATMOSPHERES

Spectroscopic criteria on the presence of an atmosphere are the most useful of all, since they identify the gases found and determine their abundance, while they do not depend on the visibility of an extended disc or on the highest qualities of atmospheric steadiness (seeing). But they are powerless to reveal the presence of such common gases as H_2, N_2 and the inert gases, and they will show other important constituents, as, for example, CO_2, CO, and N_2O only if the planet is bright enough to make it accessible to infra-red studies. Furthermore, the abundance of H_2O and O_2 in our atmosphere seriously hampers the detection of these important gases in other atmospheres, unless the objects are very bright. In that case high dispersion may be used so that planetary and terrestrial absorptions, at times partially separated by Doppler shift, can be resolved in the spectrum. Fortunately, the water vapour is concentrated in the lower layers of our atmosphere, so that it may be largely avoided if observations are made at very high altitude on days of very low absolute humidity (i.e., in practice at very low temperatures, preferably about $-40°$ C.). The smallest amount of precipitable water in a vertical column above an observatory recorded during actual observations is 0·2 mm.; but this is not good enough for future studies of Mars and Venus, and higher sites will have to be developed. But even such sites will be of no help in avoiding terrestrial oxygen. The best hope in this case is to observe the strength of the ozone bands near 3000 A. as indicators of small quantities of oxygen.

The quantitative results of spectroscopic studies derived to date, both actual amounts and upper limits (for gases where the test was negative), are collected in Table 2. Two objects mentioned in Table 1 are not listed again, Pluto and Triton. The tests to date are based on spectra of very low dispersion, covering 4000–8800 A.; they show no planetary absorptions. But somewhat better spectra may still be obtained, and their case is not regarded as closed. It is already clear, however, that at best only a small quantity of CH_4 could be present on these objects, and that they are very different from Neptune or even Titan.

The calibrations for the planet and satellite spectra are of two kinds: comparisons with the terrestrial atmosphere (useful for O_2, O_3, CO_2, H_2O), or with laboratory spectra obtained with long paths (Herzberg 1949). The last columns in Table 2 refer to the nature of the comparison. The upper limits for the O_2 and H_2O content of the Mars and Venus atmospheres are taken from Dunham's discussion (1949) of the Mount Wilson high-dispersion spectra; the other results were obtained by the writer on the basis of spectra taken at the McDonald Observatory, and calibrations obtained in the Yerkes Observatory spectroscopic laboratory. Wildt has tested the presence of formaldehyde on Venus, with negative results.

There exists a theoretical difficulty in comparing spectra observed from outside a scattering atmosphere with transmission spectra which have been used

Table 2. Atmospheric Compositions

Object	Gas	Amount (in cm. at N.T.P.)	Band used (μ)	p (Lab) (cm.)
Mercury	CO_2	<30	2·0	Atm.
Venus	CO_2	100000	Weak IR	9
	CO	<100	2·35	10·0
	N_2O	<100	2·15	75
	CH_4	<20	1·16; 1·7	4·5
	C_2H_4	<3	1·68	2·6
	C_2H_6	<1	1·20; 1·73	7·0
	NH_3	<4	1·51	1·7
	O_2	<500	0·68	Atm.
	H_2O	<10	0·72	Atm.
Mars	CO_2	440	1·57; 1·60	Atm.
	SO_2	$<0·003$	0·30	8
	O_3	$<0·05$	0·30	Atm.
	N_2O	<200	1·52; 1·68	75
	CH_4	<10	1·16; 1·7	4·5
	C_2H_4	<2	1·68	2·6
	C_2H_6	<1	1·20; 1·73	7·0
	NH_3	<2	1·51	1·7
	O_2	<250	0·68	Atm.
	H_2O	<5	0·72	Atm.
Jupiter	CH_4	15000	0·6–0·9	74
	NH_3	700	0·645	44
Saturn	CH_4	35000	0·6–0·9	74
	NH_3	200	0·645	33
	O_3	$<0·1$	0·30	Atm.
	SO_2	$<0·01$	0·30	8
Uranus	CH_4	150000	0·5–0·9	74
	O_3	$<0·1$	0·30	Atm.
	SO_2	$<0·01$	0·30	8
Neptune	CH_4	250000	0·5–0·9	74
Moon	SO_2	$<0·0003$	0·30	8
	O_3	$<0·005$	0·30	Atm.
J I–IV	CH_4	<200	0·887; 0·726	4·5
	NH_3	<40	0·786; 0·792	5·5
Titan	CH_4	20000	0·6–0·9	74
	NH_3	<300	0·645	33
Earth	N_2	625000	—	—
	O_2	168000	—	—
	H_2O	1000+	—	—
	CO_2	220	—	—
	CH_4	1·7	—	—
	N_2O	0·8	—	—
	O_3	0·3	—	—

for calibration purposes. This theoretical problem has been discussed by H. C. van de Hulst (1949). The difficulty consists of the fact that the intensity ratio of strong and weak absorption bands is larger when observed in transmission than by scattering. This is to be expected: the scattered radiation within a strong band will come from the uppermost layers of the scattering atmosphere, while the radiation in a weak band will have penetrated the atmosphere much farther, nearly as deeply as radiation in the adjacent continuous spectrum does. Consequently, the strong band will show a much smaller equivalent path length than the weak band, and is not nearly as well developed as in a transmission spectrum which matches the planet spectrum in the weak bands. This is very evident from Plates I–III*, in which planet spectra showing strong methane absorptions are compared with suitable laboratory spectra.

The quantities given in Table 2 are based on the *weak* bands, i.e. those bands in which we penetrate into a semi-infinite atmosphere nearly as deeply as we do in the adjacent continuous spectrum. This case applies to the four Jovian planets and Venus. For the other objects we see the solid surface itself, and the radiation must have passed through the atmosphere twice (for simplicity we assume that the observations were made at opposition, which is always nearly true). In a case like Mars, where the surface is seen and where the brightness of the disc is nearly uniform, the observed radiation has, on the average, gone through four Martian standard atmospheres, because the average secant for a uniformly illuminated disc is 2 (as is readily verified), while the path is traversed twice. In such cases the abundances found from the planetary spectrum must be divided by four; this has been done in Table 2. The values given are therefore *total* abundances for such objects. For the Jovian planets a factor 3 was applied instead of 4, to take account of the very strong darkening toward the planetary limb observed in the near infra-red (cf. Plate IV). (It can be shown that the factor 3 is exact for a darkening coefficient 1, as is actually found for Jupiter and Saturn at 8000 A.) The abundances so derived are of course not 'total' atmospheric abundances but refer only to the 'visible' part of the atmosphere, that part which is penetrated by the continuous spectrum adjacent to the absorption bands used. In the case of Venus no strong limb-darkening is observed, probably because we are looking down through a rather transparent atmosphere on to a layer of dense clouds. The factor 4 was applied in this case, and the abundances must refer to the part of the atmosphere situated above the clouds.

Table 2 contains only those gases for which spectroscopic tests have been possible in accessible parts of the spectrum. Very large quantities of H_2 must occur in the atmospheres of the Jovian planets, particularly Uranus and Neptune (which have much clearer atmospheres than Jupiter, so that we see much deeper); the amount may be nearly 10^2 times the measured amount of CH_4. After Herzberg discovered (1949) the quadrupole spectrum of H_2 it was hoped that the S(1) line at 8150·7 A. might be observable in Uranus. But his test has been negative, which is perhaps not surprising in view of the limited dispersion used on Uranus, 50 A/mm., 20 times smaller than Herzberg's laboratory spectra, while the expected amount of H_2 in Uranus is less than 10 times the amount used by Herzberg; furthermore, the line may well be near the flat portion of the curve of growth. With two or three times the dispersion the test might succeed. An alternative method has been proposed by Kastler (1949), which is also one of considerable practical difficulty. It proposes to use the

* All Plates are printed together, following p. 256.

Raman components of strong solar absorption lines (like H and K of Ca II) caused by the scattering on H_2. Since the computed depth of the Raman components of H and K is only about 1%, and the solar spectrum is crowded with absorption lines, only the most precise spectrophotometry may be able to show the presence of these components in the spectrum of the Jovian planets.

Four or five new absorptions, not due to CH_4 or a number of other common gases, were recently discovered in the spectra of both Uranus and Neptune (Kuiper 1949 c), in the region of 7500 A. Laboratory work on the identification is proceeding, and may lead to new clues in the interpretation of the atmospheres of the Jovian planets.

Spectrophotometric studies of the polar caps of Mars in the 1·5 μ region (made with a PbS cell) show that these caps are almost certainly composed of finely divided frozen H_2O and not CO_2 (Kuiper 1949 b). This means that gaseous H_2O must be present in the Mars atmosphere in quantities allowed by the negative spectroscopic tests. Dr. S. L. Hess of the Lowell Observatory pointed out to the writer that the Mount Wilson estimate of the upper limit of the Martian H_2O content may be somewhat too low, because of the presence of pressure effects in the terrestrial lines but the near-absence of such effects in the Martian lines. The terrestrial 'curve of growth' would therefore not apply to the Martian lines, which for a given amount of water vapour would be weaker than assumed. This problem is worthy of further laboratory work on the curve of growth of H_2O, under pressures of about 1 cm. Hg.

A word must finally be said about a curious situation existing on planets which have a hot side and a cold side, like Mercury and the Moon. Such a situation arises from a very long (or infinite) period of rotation with respect to the Sun, and the near-absence of an atmosphere. If a dense atmosphere is present, as is true for Venus, the slow rotation does not seem to matter; atmospheric convection appears adequate nearly to equalize the temperatures on the two sides, as is borne out by the radiometric measurements of temperature made at Mount Wilson. In the absence of a substantial atmosphere the temperature on the dark side becomes extremely low. For the Moon it is about 90° K. (Wesselink 1948), but for Mercury it must be much lower since the only heat supply must come by conduction through the planet (which must be negligible), to which must be added any heat due to radioactivity in the crust. In such cases, according to Watt's law, the vapour pressure over the whole planet must be that of the coldest spot. On Mercury nearly all gases must be frozen out on the cold side except H_2 and He, both of which would promptly escape from the planet's hot side (at 400° C. the mean molecular velocity of H_2 is 0·7 of the velocity of escape; for He the ratio is 0·5.) It is very puzzling therefore that observers have repeatedly reported white spots or haze on Mercury, covering up some of the surface markings. For this reason spectroscopic observations of Mercury should not be regarded as superfluous.

§4. VISUAL AND PHOTOGRAPHIC OBSERVATIONS OF PLANETARY ATMOSPHERES

This subject is beset with difficulties, both for the observer and for the astronomer who attempts to review the vast collection of published records. It is difficult for the observer because of the low resolution of his object and the inadequate picture he of necessity has of its physical conditions. This inadequate picture is expected to affect the interpretation he gives his observations, and will,

to some extent, be reflected in his drawings and written records. It is also difficult for the reviewer because he must attempt to separate the objective phenomenon from the subjective description of it. This makes it especially difficult to use observations made more than a generation ago; the reviewer needs to know much about the observer and his views so that he may understand what was meant and what was probably seen. This becomes of real importance when observations seem to contradict each other. Examples are the numerous reports of changes on the Moon, particularly reports of 'dew' or 'frost' deposits, which seem to be in flat contradiction to the (negative) evidence of a lunar atmosphere, or similar, but more serious, reports regarding Mercury on which veils have often been reported covering substantial areas of the planet. Well-known examples are the conflicting reports on the canals of Mars (e.g. Campbell 1918, Antoniadi 1910, MacDonald 1919, Evans and Maunder 1903, Pease 1924, Hargreaves 1945).

It is therefore fortunate that planetary photography has made such strides. While one may lose some definition in the smallest detail and one loses the beauty of the colour display (as well as the information contained in these colours), one does get objective records. With the technique of combining from three to ten negatives taken in rapid succession into a single composite print, the records are greatly improved in appearance and reliability. Particularly the half-tones are then brought out with great clarity and with greater reliability than any drawings could give.

In addition, photography in different regions of the spectrum has proved to be a powerful tool in the study of the planets, as the results of R. W. Wood, W. H. Wright and F. E. Ross have shown. In this manner the observer recaptures what is contained in the visual colours and, in fact, he will go considerably beyond it by extensions into the ultra-violet and infra-red (Wright 1928). This may be illustrated by a terrestrial scene, Plate V, from which the astronomer can set up many criteria for the presence of a planetary atmosphere and natural phenomena depending on the presence of an atmosphere. An intriguing problem, on which some progress has been made, is the extension of image recording to still longer wavelengths, 16,000 and 21,000 A. respectively, islands of continuous spectrum left by the terrestrial water-vapour bands. The second wavelength would show green vegetation and snow as *black*, as well as the polar caps of Mars and the Rings of Saturn. In fact, the Jovian planets themselves would be extremely dark owing to very strong CH_4 and NH_3 absorptions in that region.

The most extensive collection of *planetary photographs* in existence, well over 500,000, is that of the Lowell Observatory. A 24-inch refractor equipped with enlarging cameras is used, one corrected for blue and one for yellow light. The records extend over 45 years now, and many photographs, even early ones, are of very high quality. Only a small fraction of the photographs have been published (E. C. Slipher 1921–1940) (it would be economically impossible to publish them all); for this reason it is fortunate that the meteorological agencies of the U.S. Government are sponsoring at the Lowell Observatory 'The Project on Planetary Atmospheres' to study this vast collection of data and to determine the nature of the atmospheric circulations. The Project personnel consists of an astronomer-supervisor, two meteorologists and other members. The Project is expected to continue for several years. In 1939 the Lowell Observatory sent an expedition to Alianza, Chile, and again in 1939 to the Lamont-Hussey Observatory

at Bloemfontein, S.A., to take advantage of a favourable opposition of Mars and the excellent observing facilities there (E. C. Slipher 1940) (at the closest approaches, Mars has always a southern declination). A superb collection of photographs was obtained by this expedition. We are fortunate in being able to reproduce two of them in Plate VI. These photographs show slightly more detail than the individual drawings of Mars reproduced by Antoniadi (1930); and, of course, they are completely objective and accurate with respect to both the location and the intensity of detail. While the prevailing opinion among astronomers, that finer detail can be seen than photographed, is still correct, we must conclude that photography has superseded visual observation in several respects, owing to its objectivity, reliability and its simultaneous registration of detail on an extended disc. Visual observation, with the largest telescopes under the best conditions, will show detail 5–10 times finer than the best photographs; while in medium and small telescopes it may provide a continuous watch over the planets (e.g. cloud phenomena and sudden outbursts on the Jovian planets). On the other hand, the Lowell Observatory, in its continuous photographic coverage of the planets, has done just this for nearly half a century, as far as this is possible from a single station.

Excellent photographs have also been taken with the large reflectors of the Mount Wilson Observatory, ever since the installation of the 60-inch telescope in 1909 (by Hale, Barnard, Wood, Pease, Ross, Dunham). Mention has already been made of the important planetary photography in different colours done with the 36-inch Crossley reflector of the Lick Observatory (Wright 1924–39). Fine photographs have been taken also with the 36-inch Lick refractor since 1890 (by Campbell, Trumpler, Jeffers, Herbig). The series of Mars taken in 1909 with the 40-inch Yerkes refractor have been widely reproduced in textbooks (Barnard 1911). Superb results have been obtained at the Pic-du-Midi Observatory in the Pyrenees with a 15-inch and a 24-inch refractor (Lyot *et al.* 1943–46). Mention should be made of an early series of Mars photographs (1888 and 1890) by Harvard observers, made during an expedition to Mount Wilson (Pickering 1905, 1925 a).

Several series of *visual observations* remain of importance since they were made by the best visual observers of their period, and they cover together an interval of a century or more. The chief results on the surface markings of Mercury, and the rotation derived therefrom, were derived visually, notably by Schiaparelli and Antoniadi. Reference is made to Antoniadi's synopsis (1934) of these observations. Visual observations of Venus are extremely difficult because of the indistinct nature of the surface markings—if they can be seen at all. From observation of these low-contrast regions Schiaparelli suspected the rotation period of Venus to be equal to the orbital period, 225 days. Recently Danjon (1943) from his own observations came to the same conclusion; but as is discussed below, a shorter period is nevertheless more probable.

The most complete and reliable collection and discussion of visual observations of Mars is probably found in Antoniadi's monograph (1930). In it the distinguished observer has assembled his own life work on the planet and has added a critical evaluation of all earlier visual observations, beginning with Huygens' drawing of 1659. These earlier observations are of importance for a study of major, non-seasonal changes in the Martian topography; the recent ones suffice to define the seasona cycle. A few additional observations by Antoniadi

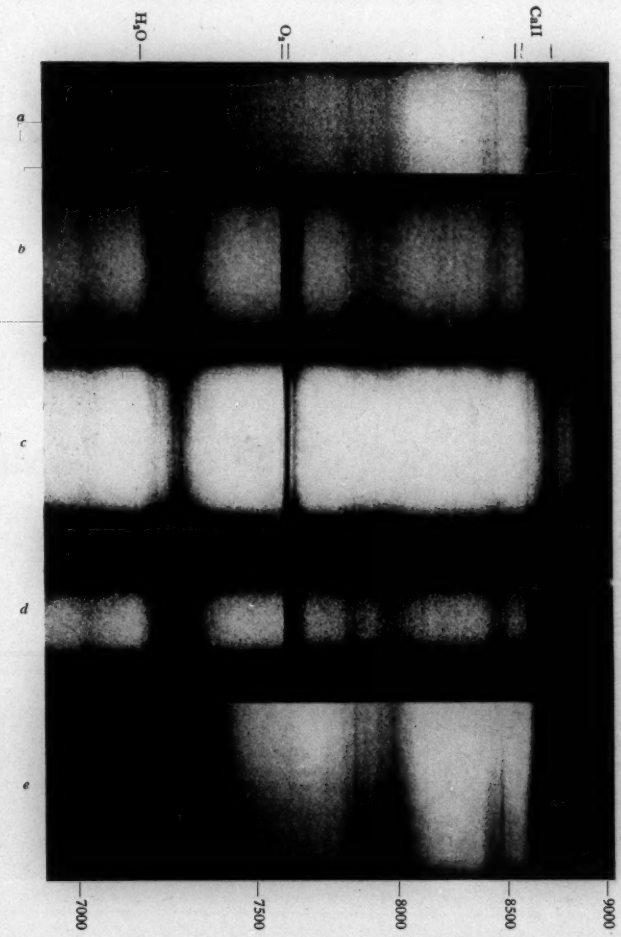
(1933–41) and the observations by Lyot and his collaborators made on the Pic-du-Midi (Lyot *et al.* 1943 b, Dollfus 1946, 1948 a) contain supplementary information. The period of growth in the knowledge and interpretation of Martian surface markings is well covered by the Reports of the Mars Section of the British Astronomical Association, in particular the reports of the oppositions of 1896 to 1916, during which Antoniadi was director of the Mars Section.

Lest the reader, comparing, for example, Antoniadi's drawings with the Lowell Observatory photograph of Plate VI, should be impressed by their similarity and conclude that average planetary drawings are as useful as photographs, reference is made to a few other sets of drawings (Pickering 1921 a, 1926 a), also made by experienced observers. It is seen that these sets would have been difficult to interpret had no other records been available. In this connection a set of six drawings is instructive, made with apertures from 8 to 100 inches (Pickering 1926 b) and showing the same region as Plate VI; the large differences in presentation need no comment. Attention is finally called to certain experiments carried out by Cerulli (1898) and other Mars observers (Flammarion 1909, Pickering 1914). They drew the Moon as seen through opera glasses which gave about the same angular disc as Mars. The 'canals' they saw and the explanation for them (a complete check was possible in this case!) are a curious commentary on the limitations of visual observation and representation. The best *direct* evidence on the nature of the 'canals' of Mars is contained in a short note by Dr. Pease (1924), who was the veteran visual observer at the 100-inch telescope.

Visual observation of the Jovian planets has a charm not dissimilar to that of our own Arctic regions. In a large reflector the colours seen on Jupiter and Saturn are unbelievably beautiful and delicate, and they foreshadow problems that it may take generations to solve. The faithful observers of the British Astronomical Association have compiled the only near-continuous record that has been published (British Astronomical Association Memoirs 1891–1943). In this connection the unique merits of the late Rev. T. E. R. Phillips should be recorded, who, as director of the Jupiter Section of the British Astronomical Association, issued the reports covering all oppositions from 1901 to 1933.

Reference must be made to *photometric* (Russell 1916, Danjon 1936, 1948, Rougier 1937) and *polarimetric* (Lyot 1929–30, Dollfus 1948 b) observations, nearly all made visually. The photometric data supply for each object the *albedo* (overall reflectivity for visual sunlight) and the *phase function* (the dependence of the reflectivity on the angle between the incident and the reflected radiation). A low albedo and a steep phase function, as is, for example, found for the Moon and Mercury, go with the complete or near absence of an atmosphere, and vice versa. As such the photometric data are contributory evidence on planetary atmospheres. The criterion may be improved by measurement of the reflection curve between 3000 A. and the blue; such work is now in progress, and should give upper limits to the *total* molecular content of objects having tenuous or no atmospheres (Titan, Pluto, Triton and the Jupiter satellites).

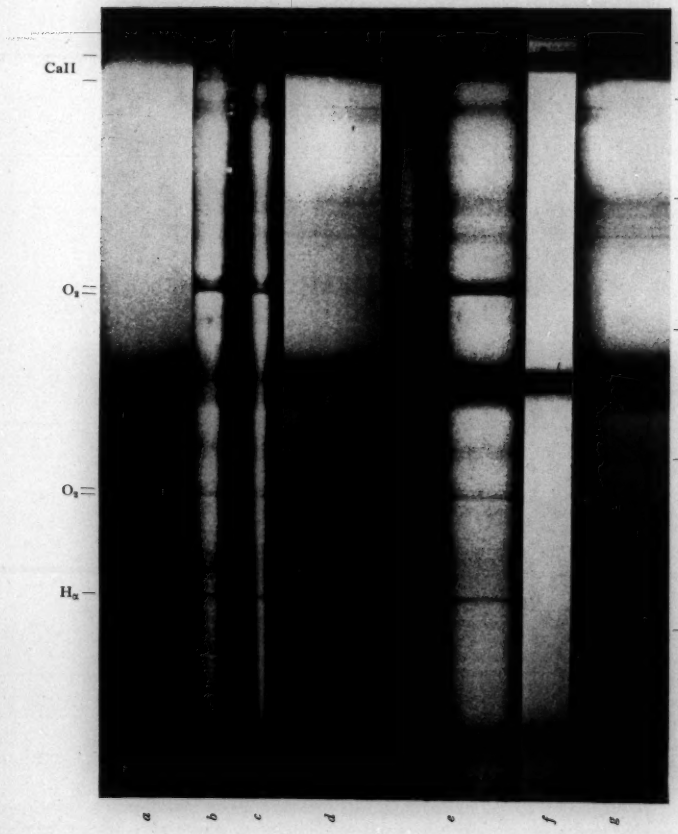
The polarimetric observations have been of very great value in the studies of Mercury, Venus, Mars and the Moon, as we shall see below, but for the outer planets the phase angles are always so small that little information may be secured (at opposition the polarization is zero).



Jupiter, Saturn, and methane spectra, 7000–9000 A. *a*, CH_4, 1 atm., 264 metres; *b*, Jupiter; *c*, Jupiter; *d*, Saturn and Ring; *e*, CH_4, 1 atm., 704 metres.

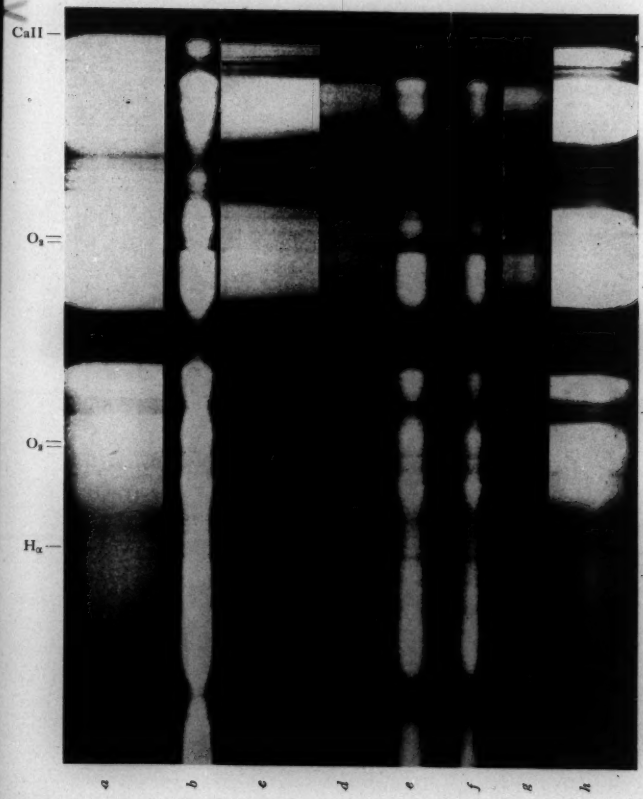
PLATE I.

To face page 256



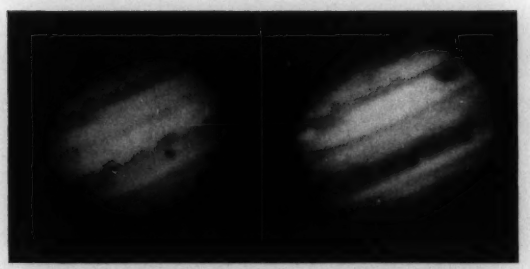
Titan, Saturn and methane spectra, 6000–9000 A. *a*, CH_4, 1 atm., 88 metres; *b*, Titan; *c*, Titan; *d*, CH_4, 1 atm., 264 metres; *e*, Saturn; *f*, CH_4, 1 atm., 704 metres; *g*, same.

PLATE II.



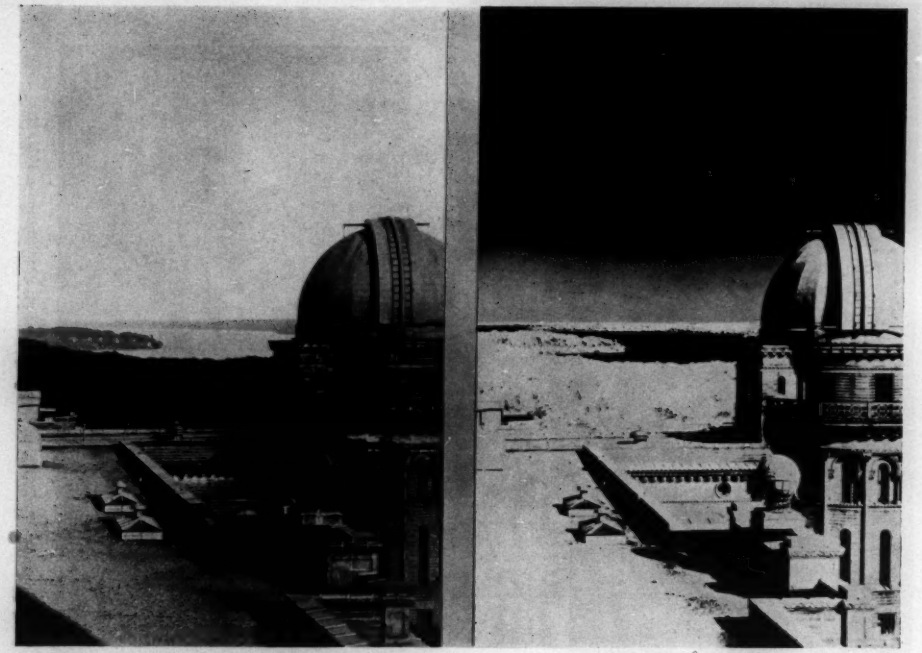
Uranus, Neptune and methane spectra, 6000–9000 A.. *a*, CH_4, 1 atm., 2,112 metres; *b*, Uranus; *c*, CH_4, 1 atm., 2,112 metres; *d*, same; *e*, Uranus; *f*, Neptune; *g*, CH_4, 1·55 atm., 3520 metres; *h*, same.

PLATE III.



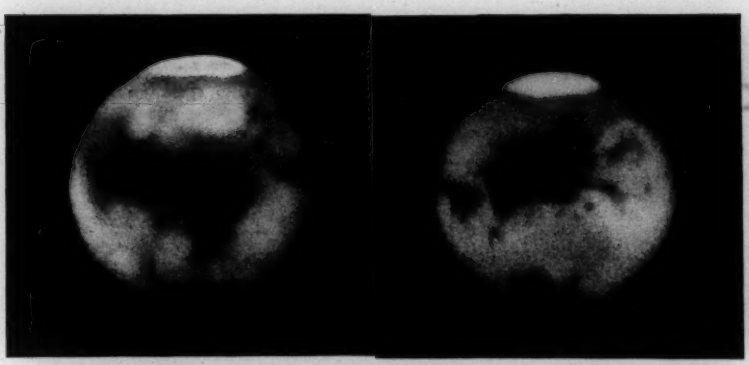
Jupiter, at 8200 A. (left) and 3500 A. (right). Dark spot near centre of disc is shadow of satellite I. Satellite itself shows bright in infra-red, dark in ultra-violet, since it is redder than Jupiter. Red spot at right, upper corner. Infra-red image printed with low contrast, ultra-violet image normal contrast.

PLATE IV.



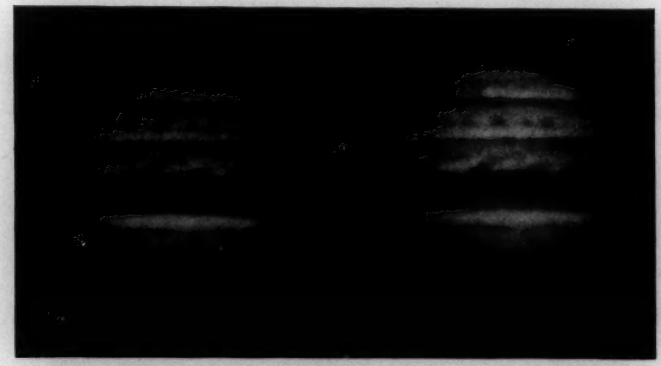
Terrestrial scene, left 3800 A., right 8500 A., showing small dome of Yerkes Observatory and Lake Geneva.
Trees are bright in infra-red; water, sky and shadows very dark.

PLATE V.



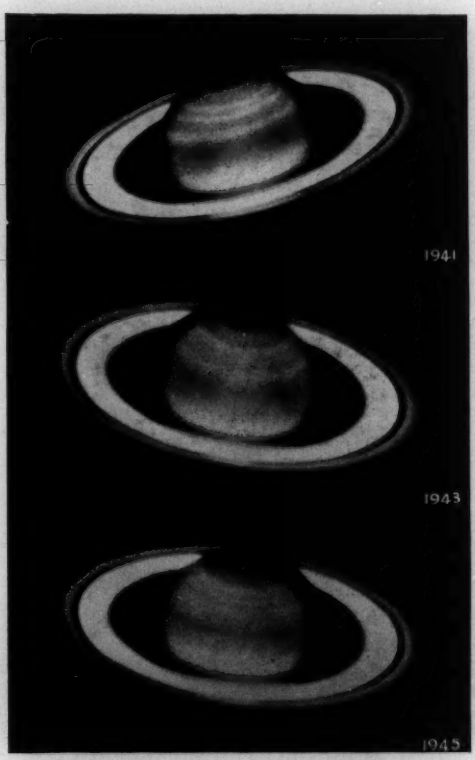
Mars, taken by E. C. Slipher with 27-inch Lamont-Hussey telescope, Blocmfontein, S.A. Right view 27th July, left view 9th August 1939; longitudes of centre, 50° and 280°. Left image is a composite print showing improved representation of half tones but slightly reduced resolution.

PLATE VI.



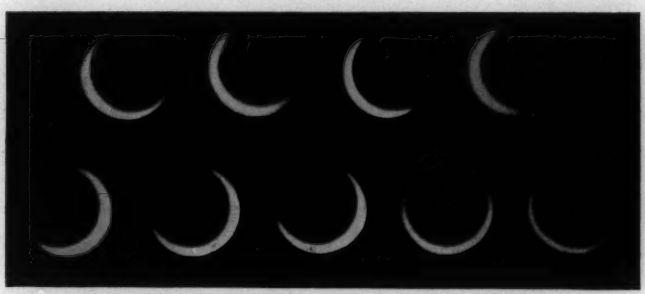
Jupiter, by E. C. Slipher, Lowell Observatory, 22nd September 1928.

PLATE VII.



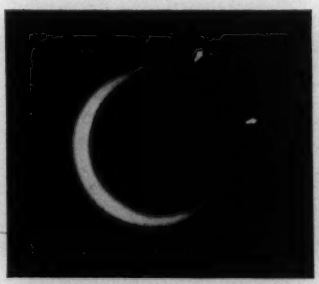
Saturn, by E. C. Slipher, Lowell Observatory.

PLATE VIII.



Venus passing inferior conjunction, June 1940, on nine consecutive days. Note extensions of cusps. Chronological order 1, 2, 3, 4, 5, etc.

PLATE IX.



Composite print of one phase, showing high clouds in Venus atmosphere.

PLATE X.

§5. OBSERVATIONAL RESULTS AND INTERPRETATION

We shall now discuss the planets in somewhat greater detail.

Mercury

Direct spectroscopic tests for the presence of an atmosphere have been negative. The visual albedo of this planet is slightly less than that of the Moon, 0·063. If we assume that less than one-sixth of this amount is due to a hypothetical atmosphere we find 0·04 terrestrial atmosphere, as the upper limit, based on the assumption of Rayleigh scattering. This test may be improved by a determination of the albedo near 3000 A. which, however, will be a difficult observation. Another determination of the upper limit may come from a photometric study of the planet very close to inferior conjunction. At that phase the hypothetical atmosphere will make its greatest relative contribution to the measured intensity. No such measures exist at present, however. The earlier measures are quite rough (Hoffmeister 1947), and the recent very fine series by Danjon (1948) includes phase angles from 3° to 123° only (the planet was invisible in the photometer near inferior conjunction). From a comparison with Rougier's excellent curve of the Moon (1933) Danjon finds agreement with Mercury within 0·10 mag. for the phase angles covered. He finds a colour index for Mercury of +1·00 mag. as compared to +1·10 mag. for the Moon; for the albedos he finds 0·063 and 0·073.

No extensions of the horns of Mercury are seen near inferior conjunction, nor is the planet surrounded by a luminous ring when situated just off the edge of the solar disc. If we assume the double horizontal refraction to be less than 6″, equal to the radius of the planet at inferior conjunction, it follows that the upper limit to the Mercury atmosphere is about 0·01 terrestrial atmosphere. We note also that in 1937 Lyot (1938) observed Mercury as a black spot on the solar corona without noticing a bright ring around the planet attributable to refraction or scattering in its atmosphere. The distance of Mercury to the solar limb varied between 1′ and 2·6′ during these observations. Furthermore, the polarization curve derived by Lyot (1929 a, b, 1930) is practically identical with that of the Moon, suggesting a similar surface texture and a small contribution (if any) from the atmosphere.

On page 253 we called attention to the haze reported by visual observers of Mercury. Since these observations can only be explained if the presence of a tenuous atmosphere is accepted, we summarize the principal conclusions of Antoniadi (1934), one of the most able and experienced of the visual observers, equipped with the largest refractor (33 in.) of Europe:

(*a*) the haze on Mercury is whitish; it occurs more frequently and is denser than that on Mars;

(*b*) it is rare that a dark surface marking on Mercury retains its normal intensity for as long as several weeks;

(*c*) the haze on Mercury presents all degrees of concentration, from very tenuous to a density sufficient to obliterate the darkest surface markings;

(*d*) the haze is usually invisible in the central portions of the disc but appears chiefly toward the limb (a result of perspective) and less often near the terminator; near the limb it may extend over an arc of 5,000 km. (3,000 miles);

(*e*) the changes observed in the haze may be very rapid; in a single day dark surface markings, 3,500 km. in size, may disappear completely, while the opposite may also occur;

(*f*) the invisibility of surface markings may last many days;

(*g*) light haze may cover a region for weeks, with improved and diminished visibility alternating;

(*h*) the north-polar region, usually clear, once showed a bright haze for six days;

(*i*) one particular dark region, on the equator and at 60° longitude from the sub-solar point (to the right in the image, with S. on top), is much more often covered by haze than any other surface marking; two regions at the same longitude but at $+40°$ and $-40°$ latitude respectively are covered less often than the equatorial region mentioned, while the whole left half of the visible hemisphere is covered still less frequently.

In a subsequent paper Antoniadi (1935 b) gives the fraction of the time during which different areas of the disc are seen covered by haze or clouds. The percentages vary from 5 to 52 for different areas. The preference of haze for the western half of the disc is noted, a fact discovered by Schiaparelli. It is tempting to attribute this to the fact that this side is leading in the orbital motion and would therefore sweep up more particles with highly elliptical orbits (cometary debris) than the following side. In this manner a very tenuous atmosphere, possibly composed of CO and inert gases, might become visible, filled with dust due to impacts, and possibly even replenished by these impacts.

The Mount Wilson temperature measurements (Pettit and Nicholson 1923, 1925, 1935, 1936) show the sunlit side to reach about 610° K. at the sub-solar point (685° K. at perihelion, 550° K. at aphelion). The dark hemisphere must be intensely cold (possibly 10°–20° K.) and tends to freeze out all but the most volatile components of the atmosphere.

Venus

The existence of an atmosphere on Venus was a certainty even before CO_2 was discovered in it at the Mount Wilson Observatory (Adams and Dunham 1932). The very high albedo, the absence of topographic detail, the extension of the horns of the crescent into a complete ring seen near inferior conjunction, and the polarization curve by Lyot, are all independent proofs. The extension of the horns has been photographed by E. C. Slipher at the Lowell Observatory in 1938 and by Edson *et al.* on an expedition to Table Mountain, California, in 1940. This expedition was privately organized by four students; the results are summarized in the Appendix. The Lowell series consisted of 500 photographs, two of which have been reproduced (E. C. Slipher 1939, Whipple 1941). Both the Lowell and the Edson photographs show local intensifications of the twilight ring, indicating differences in cloud level over regions having dimensions no less than 20°–30°.

The ultra-violet photographs of Ross (1928), and those by Quénisset (1911, 1921) and Wright (1927 a, 1928), show bright and darker atmospheric markings of dimensions comparable to those suggested by the intensifications in the twilight arc. Ross found the 'dark' markings to be about 24% fainter than the bright portions of the disc; in blue light the contrast fell to 4%. In visual light these (or similar) markings have extremely low contrast which explains the difficulty referred to on p. 255 of establishing the period of rotation. It is assumed that

the bright portions of the ultra-violet image represent layers of haze or fine dust in the upper atmosphere, with the darker parts representing regions of greater penetration. The uniformity of the visual and infra-red images (up to 1μ) must mean that in these wavelengths we are looking down on a dense haze or cloud layer, nearly opaque up to at least 1μ. Nevertheless, in visual light the scattering properties of the Venus atmosphere are not homogeneous with respect to the distance from the terminator (the edge of the sunlit side), as was shown by Minnaert (1946) from an interesting application of the Helmholtz reciprocity principle in optics.

Ross (1928) has used his photographs, together with all other pertinent data, in a careful discussion of the period of rotation of the planet. He rejects the long period of rotation, 225 days, suggested by Schiaparelli (and later by Danjon 1943) on three grounds: (*a*) if the same side of the planet were always exposed to the sun, a stationary flow pattern in the atmosphere would be expected; the ultra-violet photographs show markings changing from day to day without a regular pattern; (*b*) the dark markings show a banded structure, at right angles to the terminator, suggesting fairly rapid rotation ($P<20$ days, say); (*c*) it is difficult to conceive how the small temperature difference between the sunlit and the dark side, observed radiometrically, could be maintained by convection alone. On the other hand the spectroscopic determinations of the rotational velocity made at the Lowell and Mount Wilson Observatories preclude a direct rotation with a period much shorter than 20 days (a retrograde rotation is not expected, but the obliquity may be as high as 90°; these conclusions follow from an unpublished theoretical study by the writer). Ross favours a rotational period of 20–30 days, as a compromise. Clearly, this problem is still unresolved.

The vertical structure of the Venus atmosphere was studied many years ago by Russell (1899) on the basis of the observed extensions of the horns or cusps, shown near inferior conjunction. Russell found that a strip 1·2° wide is illuminated by the Sun outside the geometrical sunlit hemisphere, to an intensity high enough to be seen against our own daylight sky. Fainter extensions, bluish in colour, have been reported by Schroeter from night-time observations (more data are required on the latter). Russell shows that the 1·2° arc is not due to *refraction*; a refracted solar image would show first on the far side of the planet, not as an extension of the cusp.* A truly refracted image is seen when the planet is only a few minutes of arc from the Sun; it is much brighter than the normal Venus image. Russell concludes that a haze layer about 4,000 ft. high penetrated by sunlight will account for the observed extension of the cusps. Above that layer Russell assumes the atmosphere to be relatively transparent, with a horizontal refraction of less than 12′. The 1,000-metre atmosphere of CO_2 found spectroscopically would cause a horizontal refraction of about 6′; but this amount of CO_2 must include most of the haze layer itself. The 'ultra-violet' clouds shown in Ross's photographs must float above the haze layer, possibly as fine extensions of it. The bottom of the haze layer may, according to Russell, be either the planetary surface or an opaque cloud or fog layer. It is now assumed that the second alternative applies; for, if the haze layer is transparent enough for sunlight to penetrate two double horizontal paths through it, as Russell's derivation

* The criticism against this conclusion by E. Schoenberg (*Astr. Nachr.*, 1949, **277**, 123) is invalid. Atmospheric dispersion, invoked by Schoenberg, is much less important than the change of horizontal refraction with altitude in the atmosphere; the latter effect was properly considered by Russell.

requires, it should certainly be possible to see through it vertically. Since no 'brightening toward the limb' is found (the opposite is true), and the planetary surface is unlikely to have the albedo 0·7 or higher, we are led to assume that we are looking down through the haze on an even denser layer. This layer appears slightly yellowish, as has been pointed out by Ross and others, and must be responsible for the colour index of 1·00 mag. (see below).

The Lowell observers have reported remarkable fluctuations in the emitted planetary radiation and have attributed this to changes in altitude in the radiating layer (Lampland 1941, Adel 1941). The Mt. Wilson observers have not noticed such changes (Pettit, personal communication). If the changes are real they imply that haze particles larger than 2μ (i.e. $11\mu/2\pi$) stop the penetration in the 11μ region, and that the effective upper boundary of these particles undergoes large fluctuations in altitude. The writer found (1950, unpublished) that the CO_2 content measured spectroscopically at 8700 A. undergoes large changes, partly rapid and erratic and partly correlated with phase. The crescent at inferior conjunction, instead of showing more CO_2, as would be expected on the basis of a stratified atmosphere, showed actually less than the average of the full disc. This result can be understood if the particles are rather well mixed with the gas up to very great altitudes and have a strong forward scattering (a common property of particles several times larger than the wavelength of light).

The scattering properties of the particles are best studied from photometric data. These data are of two kinds: (*a*) total intensities of the reflected light of the planet as a function of phase angle, and (*b*) distribution of light over the planetary disc for a given phase angle. The second problem has hardly been touched (Ross 1928) because of its difficulty. The first has been attempted by Gerasimovič (1937) on the basis of Müller's measures. Greatly improved empirical data are now available (Danjon 1948), for phase angles from 1° to 171°, and also the theory of radiative transfer in scattering atmospheres has been greatly advanced by the work of Chandrasekhar. A discussion of Danjon's measures on the basis of Chandrasekhar's theory is currently being made by H. Horak.

The radiometric measures show remarkably little difference between the sunlit side and the dark side (Pettit and Nicholson 1924 b, Coblenz and Lampland 1924 b). The temperatures derived are somewhat discordant, the Lowell observers finding 50°–60° C. for the bright side (Menzel, Coblentz and Lampland 1926), and the Mt. Wilson observers about −32° C. for both the bright side and the dark side (Pettit and Nicholson 1924 b, 1930). The discordance might be due to variability of the cloud layer. These temperatures refer to atmospheric layers, in view of the powerful absorption and emission of CO_2 in the 8–15 μ band and the presence of large particles. This is also suggested by the variability in emission referred to above. Adel (1937), from the intensity distribution within a CO_2 band, derived an atmospheric temperature in excess of 50° C. for the CO_2 on the sunlit side; and Wildt (1940 a) computed 100°–120° C. for the planetary surface below. The surface temperature on the dark side, while higher than the atmospheric temperature measured, may not be much higher—since the convection on Venus will lift the particles much higher on the sunlit than on the dark side (Ross 1928). Ross discusses some evidence showing that the haze layer near the terminator is sloping toward the dark side.

Danjon (1948) found no noticeable deviations of the planet's brightness from a mean light curve, consistent with the absence of spots in visual light. The colour index came out at +1·00 mag., rather yellower than usually suspected.

Further information on the haze layer is derived from the polarization curve. Lyot has developed a planetary polarimeter of unprecedented precision (about 1 part in 1,000, some 20 times better than earlier astronomical equipment of this kind) and obtained a most interesting curve (1929 a) showing the polarization as a function of the phase (the angle between the directions of Sun and Earth as measured at the planet). He showed that the polarization was probably produced by multiple scattering (i.e. scattering by a dense cloud of particles each having a high albedo), and that it had some resemblance to scattering by terrestrial haze consisting of water droplets about 2 μ in diameter. In view of the fairly high daytime temperature measured on Venus and the absence of water vapour from the spectrum it is difficult to accept that water droplets are present. Van de Hulst showed (1949) that a combination of a larger refractive index (say 1·5 or 1·6) and larger droplets (say 4–5 μ) would give a similar polarization curve as water droplets of 2 μ. Measurements of polarization curves at 1·6 and 2·1 μ are now in progress; they, together with Lyot's visual curve, should define both particle size and refractive index. It may be possible also, in future work, to distinguish between spherical particles (droplets) and crystals or dust. This distinction is perhaps the most important of all Venus problems. If the refractive index and the shape of the particles indicate that the haze is in fact dust (as was assumed by Lowell and others many years ago), it could well be attributed to the violent convection from the hot to the cold side of the planet.

Reference is made to two papers by Wildt (1940 b, 1942) on the chemistry of the Venus atmosphere. The apparent absence of water from Venus and its presence on the Earth pose an interesting problem, the solution of which will tell much about the origin of the terrestrial planets. This absence and that of oxygen, with CO_2 present in quantity, suggest that life has not developed on this planet.

Mars

The existence of a Martian atmosphere was established from observations of the polar caps and their seasonal variations, the formation and the disappearance of clouds, the dense haze in which the planet is enveloped if seen in blue or violet light, the existence of a twilight arc, the sudden variations in the polarization curve, and the seasonal variations in colour and intensity of the dark markings on the disc. These markings cover about one-third of the surface area. CO_2 on Mars was discovered at the McDonald Observatory in 1947. It was also shown that the polar caps consist of H_2O frost, so that some H_2O vapour must be present in the Martian atmosphere. All other spectroscopic tests applied so far have been negative (cf. Table 2).

The albedo of Mars, while not as low as that of the Moon or of Mercury, is not as high as that of the partially cloud-covered Earth, and much less than that of Venus and the Jovian planets. This is consistent with the fact that the Martian surface features are easily seen (except during occasional periods when there seem to be large 'dust storms' or veils of haze). It indicates an atmosphere less dense than our own.

Of all the planets and satellites Mars resembles the Earth most closely, not in mass and dimensions but in appearance. For this reason it has been the object of very careful study and much speculation. The period of rotation is 40 min. longer than that for the Earth; the axis of rotation is tipped 25° with respect to the normal to the orbital plane, as against 23½° for the Earth; and the year is nearly two of our years. Seasonal variations are therefore very similar to those on Earth, although some important differences exist:

(*a*) Mars has no seas, nor even small bodies of liquid water; this follows from direct observation (cf. Plate V for criteria) and from the very low vapour pressure found in the atmosphere; it is also consistent with the low mean temperature and the large temperature range on the planet (see below);

(*b*) Mars' greater average distance to the Sun (1·523 astr. units) implies a black-body equilibrium temperature of $1{\cdot}523^{-1/2} = 0{\cdot}81$ of our own. This value is somewhat increased if the known deviations from black-body absorption are considered: the albedos of Mars and the Earth being 0·15 and 0·29, the absorptivities are 0·85 and 0·71. Therefore the mean Martian absolute temperature should be $(0{\cdot}85/0{\cdot}71)^{1/4}$ times 0·81, or 0·89 times our own, or 30° c. lower.

(*c*) The lower density and particularly the very low water-vapour content of the Martian atmosphere cause its 'greenhouse effect' to be much less than our own. For a large area on the Earth the layer radiating freely to space is not the ground but the top of a layer of clouds or humid air with temperatures between 0° and −20° c. The ground temperature may then be much higher than that of the emitting layer and there will be little difference between the day and night temperatures on the ground. When the ground temperature of Mars is near its maximum value, about 0° c. at the planet's average distance to the Sun, most of its temperature radiation will escape freely to space, since the atmosphere will probably offer little obstruction to $\lambda < 14\,\mu$. For longer wavelengths, however, CO_2, and to some extent H_2O, will absorb powerfully, so that, as the night temperature drops, the net emission will shift more and more to the atmosphere. While the Martian daily temperature range is therefore expected to be very large by our standards, it is nevertheless moderated by the atmosphere, and not nearly as large as, for example, on the Moon. The daily range in the equatorial regions may be from 0° at noon to −60° or −80° c. at night. This conclusion is consistent with radiometric observations made at the Mount Wilson and Lowell Observatories.

(*d*) The Mars orbit has an eccentricity of 0·093. The percentage range in distance to the Sun is therefore twice this amount, the range in radiation received per unit area four times, and the range in equilibrium temperature equal to the amount itself. The annual variation due to this cause is therefore about 25° c. This temperature difference has important effects on Martian meteorology. It causes the summer of the southern hemisphere to be *warm* (but it is also *short* because of Kepler's law of areas), while its winter is long and cold. The opposite holds for the northern hemisphere. Yellow clouds (see below) are most prevalent at perihelion, white clouds at aphelion (Antoniadi 1930).

The seasonal variations of the polar caps are well described and illustrated (60 drawings) by Barnard (1895) (cf. also Pickering (1925 b), Antoniadi (1930) and E. C. Slipher (1938)). Wright showed (1924–27) that the polar caps have associated with them a cap of atmospheric haze, which looks like a concentration in the general haze observed in violet light. The cap visible on the ground

consists of a low-temperature H_2O frost deposit, similar to the one produced on a piece of solid CO_2 exposed to humid air (Kuiper 1949 b). The albedo of the cap is in the Martian Spring only about twice that of the rest of the planet, or 0·3–0·4, indicating a lessening of albedo with age, as is true for terrestrial snow.

Clouds on Mars are much scarcer than on Earth. They are of two kinds, low-lying *yellow* clouds, possibly composed of dust, which sometimes cover large areas of the planet and obliterate their detail even in visual light; and high-lying clouds which appear to be no more than concentrations in the general white haze enveloping the planet; the latter may be compared with light cirrus clouds. The white clouds appear from 3 to 10 times as dense as the general white haze, but they are still nearly transparent to red light. They are often seen near the morning and evening limbs of the sunlit hemisphere, and sometimes protrude into the twilight zone. If this happens, the astronomer may make a rough determination of the cloud's altitude. Values are found ranging from 4–20 km. (Antoniadi 1930). E. C. Slipher observed three such clouds in 1922 and found 15,000 feet ($4\frac{1}{2}$ km.). A very striking cloud feature beyond the terminator was observed by Van Biesbroeck (1924); it indicated an extensive nightly cloud cover dissolving at sunrise.

Ordinarily white clouds are seen forming less than one hour before sunset and are seen to dissolve less than one hour after sunrise. This suggests a temperature effect except that it is surprising that the atmosphere would follow the ground temperature so quickly (contrary to the Earth). A large part of the increased visibility of clouds towards the limb may be due to perspective. Sometimes extensive haze fields or white clouds are seen throughout the Martian day (E. C. Slipher, 1921 a, 1922, Pickering 1925 a, Antoniadi 1930), while occasionally a cloud is even seen to form at Martian noon (Wright 1927 d). The cloud observed by Wright was forming on four consecutive days in October 1926 in the same spot, lasted each time until after sunset and was gone the next morning. Wright assumed it to be caused by vertical convection due to the sun's heat on the planetary surface.

It may be hoped that the crystal shape of these condensation products will be derivable from polarization measurements. Such work is in progress at the Pic du Midi (Dollfus 1948 b).

Wright (1924–39), and also Ross (1926), have demonstrated with beautiful photographs that the Martian atmosphere in violet light is quite opaque, while being nearly transparent in yellow and red light. Similar, less explicit results had been derived by Pickering (1905, 1925 a, c), de la Baume Pluvinel and Baldet (1909), Tikhov (1910) and the Lowell observers. While it is true that normally the Martian surface detail cannot be seen by $\lambda < 4500$ A., there are periods when, at least locally, the limit is much shorter (Pickering 1921, 1925 a, E. C. Slipher 1927, 1937, 1938), while sometimes it is longer, and the planet is enveloped in white haze even visually. The very dense atmospheres derived on the assumption that the opacity in the violet is due to Rayleigh scattering are therefore not real. The most crucial published observation in this connection is the one by E. C. Slipher in 1937 which showed Syrtis Major as a dark spot in blue light on two consecutive nights. The upper limits to the molecular content of the Martian atmosphere set by this observation is quite low (see below).

The Martian haze is white or bluish white, and owes its normal prominence in blue light partly to the greater contrast with the orange planet and partly to a

somewhat increased opacity for blue and violet light. Molecular Rayleigh scattering is not very important even near 3000 A., since the albedo does not follow the λ^{-4} law. The white haze makes the limb of the planet appear whitish visually. Wright has pointed out that the haze near the limb is 'lumpy', which is consistent with the close relationship between haze and clouds noted above. It is very probable that both consist of ice crystals. The alternative hypothesis that the condensations are solid CO_2 would require that the haze layer has an elevation of some 45 km. and a temperature of $-160°$ C. This height is much greater than that found for the white clouds and seems to rule out this hypothesis.

We can only mention in passing the observations of the twilight arc made by Lowell and others, and Lyot's discovery of sudden changes in the polarization of the planet's light, coincident with the appearance of yellow 'dust' clouds. Nor does space permit the discussion of the seasonal and irregular changes in the colour and extent of the dark surface markings often held to be vegetation. Reference is made to Antoniadi's description (1930) of the colour changes and other seasonal effects and to his and E. C. Slipher's discussion (1938) of non-seasonal changes in the surface markings. Infra-red spectrophotometric data concerning these markings were obtained by the author (1949 b), while previously spectrophotometric data in the visual range had been obtained by Tikhov (1945) and Millman (1939), as well as by V. M. Slipher, at the Lowell Observatory (1924).

The Martian 'canals' to the extent that their reality has been confirmed appear often to be composed of small irregular markings sometimes arranged in a chain, and sometimes without any dominant structure (Antoniadi 1930, Pease 1924, Dollfus 1948 a). Their colour changes appear to follow those of the larger adjacent markings.

The total atmospheric pressure at the surface of the planet has been estimated by three methods:

(*a*) The height of the clouds in the Mars atmosphere is used (Hess 1948); the result is 0·08 atmosphere, but depends on the correctness of certain meteorological assumptions as to cloud formation on Mars;

(*b*) The difference in photographic and visual albedo of the planet is used (Menzel 1926); this can give only an upper limit since the atmospheric scattering in the blue is primarily by particles and not by molecules (E. C. Slipher 1937); the true value will probably not exceed one-quarter of the upper limit (0·07 atm.) found, i.e. 0·02 atm.;*

(*c*) Dollfus (1948 b) used his measurements of polarization to derive a pressure of 0·08 atm.; on the other hand, Lyot, by the same method, had previously found 0·01 atm.

Method (*a*) might be extended to include the conditions for atmospheric support of haze and dust. Rough sizes for the particles should be derivable from the relation between reflection and wavelength, while at least the horizontal wind velocities may be estimated directly from the drift of clouds. Velocities of about 20 miles per hour are found (Antoniadi 1930, E. C. Slipher, and others). Method (*b*) should be refined by the use of the relation between albedo and wavelength from 3000–5000 A. coupled with a study of the limb brightening in these

* Dr. E. C. Slipher has shown the writer unpublished observations of 9th and 10th October 1941 in which the clearing of the Mars atmosphere in blue-violet light is even more remarkable. These observations place the upper limit of the mass of the Mars atmosphere at about 0·01 Earth atmospheres. Most of this mass is probably argon, as has been pointed out before.

wavelengths. Provisionally we adopt $p = 0{\cdot}01$ atm. (or a total atmospheric mass of 1/40 terrestrial atmosphere) as the best value.

The low humidity (cf. Table 2) puts the dew or frost point below $-50°$ c., but for various reasons it is unlikely that it falls much below $-60°$ c. The polar caps and the icy haze are therefore expected to form below that temperature ($-60°$ say), which is consistent with the radiometric observations (Menzel, Coblentz and Lampland 1926, Pettit and Nicholson 1924a, 1925, 1930). In the past it has usually been assumed (e.g. Coblentz 1925a, 1927, Pickering 1925b, 1926c) that the 'melting' of the polar cap defines the isotherm of 0° c. The inconsistency of this assumption with the observed radiometric temperatures led then to the further assumption that somehow the measured temperatures were too low, presumably caused by absorption in the overlying atmosphere, and that the smaller Martian polar caps indicate that Mars was actually warmer than the Earth. In view of the transparency of the atmosphere to red light we must assume that the general absorption near 10μ is negligible except as caused by thick clouds and by molecules like CO_2. The effect of molecular absorption is not large (probably less than 10° c.), as follows from temperature measurements made on the sunlit limb of the planet when there is a considerable phase angle, which show values only slightly less than those at the centre of the disc. From meteorological data on the rate of the evaporation of snow at very low temperatures it should be possible to make some direct estimate of the thickness of the Martian polar frost deposits; it is probably measured in inches rather than in feet. From Table 2 it follows that the upper limit to the total amount of precipitable water in the Martian atmosphere is only 0·05 mm. If this full amount were collected on one polar cap extending down to 60° latitude, the thickness of the layer of ice would be $2 \times 0{\cdot}05/(1 - \sin 60°) = \frac{3}{4}$ mm. Deposited as snow, the layer would be at most 1 cm. high. Since nearly the complete polar cap evaporates in its own summer, it appears that a good fraction of the water vapour is transported by the Martian atmosphere from pole to pole twice per Martian year; it is unlikely therefore that the total water supply on the surface of Mars is more than 3 to 10 times higher than the amount just calculated.

When one of the polar caps is rapidly being evaporated away, the adjacent dark areas show their maximum coloration (green, yellowish green, or bluish green). It is this fascinating phenomenon that has inspired astronomers and popular writers to speculate about the nature of 'life on Mars'. We cannot refer to these speculations here since their interest is largely historical. We wish, however, to call attention to a summarizing article by Coblentz (1925b) which shows that speculation by an experienced scientist may mean true progress. At present the hypothesis that the green areas represent plant life is much more probable than when Coblentz wrote his article; for CO_2 and H_2O are now known to be present, so that photosynthesis is actually possible.

There is no reason, of course, why the green areas, if actually vegetation, should contain any of the terrestrial species. Nevertheless, in view of the enormous differentiation of terrestrial plants and the range in habitat, a comparison of the properties of the Mars areas and terrestrial species is instructive. In summer the vascular plants (spermatophyta and pteridophyta) show a reflection spectrum containing the chlorophyll bands near 4000 and 6800 A., and the liquid-water bands near $1{\cdot}5\mu$. These features are absent from

the green areas on Mars (Kuiper 1949 b). It appears therefore that our vascular plants are not responsible for these areas, a conclusion that might have been anticipated from the rigours of the Martian nights. Only in the long summer day of the polar cap, after the snow is gone, might such plants be able to live. Lichens and mosses show a reflection spectrum not very different from the Martian green areas, and they could withstand the Martian nights. Certain differences exist, however, which show this identification to be too simple: (*a*) lichens do not show a seasonal colour cycle, while several of the Martian areas do; (*b*) some Martian years show a substantial increase in the green areas, amounting to hundreds of square miles: such rapid changes are not observed among lichens; (*c*) the colour cover on Mars is in the nature of a blanket: the colours are vivid, indicating little space between the elements. Lichens are usually patchy.

Mention should be made of an observation by E. C. Slipher (1923) that a green area was darker after a white cloud had covered it. The same author has observed a large region suddenly getting darker (not green), which he attributed to thawing of the soil. Additional observations of this type would be most instructive; it does not seem consistent with the water-vapour data that liquid water was present in either case. Precise photometric measurement of a given green area in its daily rotation, when Mars has an appreciable phase angle, may disclose a varying proportion of light and shadow attributable to vertical structures.

Jupiter

The atmospheric phenomena of Jupiter appear from the Earth the most vigorous of all. There are probably no two days on which Jupiter shows precisely the same features. The planet rotates in somewhat less than ten hours, but the angular velocity is not uniform. It takes the bright equatorial zone five minutes less to make one revolution than it takes the rest of the visible planet (as was discovered by Cassini in 1690); minor differences, all less than 50 seconds, exist between the other zones (Phillips 1915).

The planet shows a system of bright zones, roughly parallel to the equator, alternating with darker belts. The zones and belts undergo very considerable variations in width and detailed structure. The colours seen in them in a large telescope are among the most remarkable visible in the sky. We quote from a brief article by Pease (1925) based on his observations with the 100-inch telescope: "The following colours may be seen with the 100-inch reflector in various portions of the planet: belts—red, brown, chocolate, copper, buff, tan and fawn; light-coloured zones and polar regions—yellow, light buff, cream, light tawny, and bluish white; spots or eyes—black, red, white, and copper."

Not only are the colours intriguing but the relative motions of the different parts are a subject for endless study (British Astronomical Association Memoirs). Most or all of these details can be photographed with the larger telescopes, but even the best photographs do not convey the vivid impression of infinite detail acquired visually under the best conditions. We again quote Pease (1925): "The most fascinating details, however, do not show on the photographs, being fine threads and wisps, identical in appearance with delicate cirrus clouds, that occur in the central zones and belts. They often take the form of cusps, a dozen or more scattered along the belt, their convex sides all pointing alike. They occur with the cusp standing perpendicular to the belt, but when they extend into a zone the few sketches show them to be inclined, as though the belt and zone

had different velocities." An intimation of these statements may be had from inspection of Plate VII, which shows two photographs of Jupiter taken in 1928, kindly made available by Dr. E. C. Slipher of the Lowell Observatory.

The eccentricity of the Jupiter orbit is 0·048, so that, for reasons similar to those discussed under Mars, a 5% change in surface temperature (about 7 or 8° C.) is expected during the Jupiter year (11·86 of our years). It appears that this annual variation has observable effects on the colour of the bright equatorial zone, as was discovered by Williams (1920–1936 b). This zone has a tawny coloration for a period of three years around Jupiter's aphelion; the colour studies covered 150 years of visual observation.

The planet's obliquity of the equator is only 3° and no appreciable seasonal variations would be expected. Yet such a variation appears to exist, as was also discovered by Williams (1899, 1930, 1936 a). Each of the two dark equatorial belts undergoes a colour cycle in the course of the planet's revolution around the Sun, from quite red to very little red. The phase of the N. belt is opposite to that of the S. belt. The observations extend over about one century.

These important empirical results provide clues to the causes of the vivid colours observed on Jupiter. They were followed up by a chemical study by Wildt (1939 a) according to which metallic sodium reacting with ammonia may be held responsible for the whole range of colours observed on Jupiter. Wildt accounts for the absence of such vivid colours on Saturn by assuming the temperature on that planet to be somewhat lower.

The albedo of Jupiter is high, as is to be expected for a dense atmosphere.

No permanent features are visible on the planet, indicating that the atmosphere is deep, say of the order of 100 km. The object coming nearest to a permanent feature is the great Red Spot, which has been traced back for nearly three centuries (since Hooke's observation of 1664), the full period of telescopic history. However, its angular motion is not constant. The total amplitude with respect to a uniformly moving point having the same average motion is 850°, as was derived by Peek (1939). The most reasonable interpretation given to the Red Spot is that it consists of an atmospheric cap hovering over a large solid mass floating at a deeper level. The changes in angular velocity can then readily be explained as due to small vertical displacements of this floating mass (since angular momentum will be preserved (Peek 1939)).

Another region of special interest in the same latitude belt is the South Tropical Disturbance. The ultra-violet photograph in Plate IV shows the Red Spot in the upper right-hand corner and the South Tropical Disturbance near the left edge of the same belt. When smaller spots approach the preceding end of the Disturbance they come up to it and then appear to turn back, as if they were reflected. This phenomenon was discovered by Peek (1940), who illustrated it on a very instructive chart. The Disturbance rotates faster than the Red Spot and has been seen to overtake it six times between 1902 and 1914 (Phillips 1915). The Disturbance is attracted by the Red Spot; it is accelerated on approach and retarded after passage. Phillips (1915) interprets this as being caused by a vortex property associated with the Red Spot. The Red Spot itself is accelerated during the passage of the Disturbance. Many other remarkable phenomena are described in Phillips' report (1915) and in the reports of the Jupiter Section of the British Astronomical Association (B.A.A. Memoirs). Reference is made also to a paper by Hargreaves, Peek and Phillips (1928), entitled 'The Revival of Activity in the Southern Hemisphere of Jupiter', and

comments on this paper based on photographic observations at the Lick Observatory by Wright (1929). We quote Phillips (1915) on the behaviour of belts: "A whole belt may fade away almost to invisibility or return from such a condition to a state of great strength in the course of a few weeks." A good illustration is given by E. C. Slipher (1938). The fact that the principal Jovian phenomena are arranged in zones and belts parallel to the equator is interpreted by Lau and Phillips as due to the very rapid rotation. A theoretical study of the differential rotation of the equatorial zone is contained in a recent paper by Schoenberg (1948).

The relative motion of separate spots often attains 200 miles per hour. Spots moving with such a velocity were photographed by E. C. Slipher in 1928; a reproduction may be found on p. 157 of Whipple's book (1941).

The Jupiter photographs taken in light of different colours by Wood (1916) and Wright (1926, 1927 c, 1927 d, 1928, 1929) show a small limb-darkening in the ultra-violet on the equator, but a strong darkening towards the poles; the infra-red and red images show a uniformly strong darkening all around the limb. Plate IV illustrates these findings. The writer found the following coefficients of limb-darkening: equator at 3500 A., 0·48; equator at 8500 A., 1·0; planet as a whole, at 1·3 and 1·6 μ, 1·0. The data for $\lambda > 1\,\mu$ were obtained from drift curves taken with an infra-red spectrometer. Van de Hulst (1949) has discussed the implications of these measures for the scattering properties of the Jupiter atmosphere.

Direct temperature measurements show that little if any heat derives from internal sources of the planet (Menzel, Coblentz and Lampland 1926). Peek (1937) studied 'The Physical State of Jupiter's Atmosphere'. From his data (p. 581) it would appear that the Jupiter atmosphere might be about 200 km. deep. The time is definitely past when one might freely speculate on a Jupiter atmosphere 10,000 miles in depth. Wildt and Peek have shown that severe restrictions are imposed by the known low density of the outer mantle of the planet. However, no precise data on the density gradient and the atmospheric depth can be derived until more is known about the thermal properties of hydrogen under moderately high pressures.

Saturn

This planet has an average density of only 0·71 C.G.S. and presumably a composition similar to Jupiter, the smaller density being caused by reduced compression owing to the smaller mass. Saturn is nearly twice as far from the Sun as Jupiter (cf. Table 1) so that we would expect its absolute surface temperature to be $\sqrt{(5{\cdot}20/9{\cdot}54)} = 0{\cdot}74$ of Jupiter. It is at first surprising, therefore, that the radiometric measures give 130° K. for Jupiter and 128° K. for Saturn (Menzel, Coblentz and Lampland 1926). The first value is expected from the distance to the Sun, the second value is about 30° higher. The high measured value for Saturn is probably due to a penetration into the atmosphere, in the 8–14 μ band, well below the average radiating layer of the planet; this average layer is determined by the atmospheric absorptions in the 20–60 μ region, around the maximum of the Planck curve. Owing to the very strong double band of NH_3 near 10·5 μ (extending from 9–13 μ) and the abundance of NH_3 in Jupiter, the 8–14 μ radiation in the Jupiter atmosphere is expected to correspond to the outermost layers (the 7–9 μ region is filled with the ν_4 band of CH_4, even more abundant). But on Saturn the NH_3 content observed in the red light is reduced in spite of the fact that the layer so penetrated is far deeper, as is shown by the

increased CH_4 content (Table 2). If the radiation between 9–13 μ comes essentially from the top of the NH_3 layer for both Jupiter and Saturn, the equal radiometric temperatures of the two planets is explained. This is seen as follows: 1 cm. atm. of NH_3 is nearly opaque between 9 and 13 μ (Hsi-Yin Sheng *et al.* 1941). At $-150°$ C. $= 123°$ K. the vapour pressure of NH_3 is 0·00056 mm., so that a path length of 13·5 km. of saturated atmosphere at this temperature will form an effective radiating layer between 9 and 13 μ. (Near-saturation is expected because of the great quantities of NH_3 that must be present in the deeper layers of all the Jovian planets.) Because of the very large differential motion between the different latitude zones the atmosphere is expected to be turbulent. Therefore the temperature gradient in the outer layers must approach the adiabatic one. Since H_2 must be by far the most abundant constituent of the atmospheres of the Jovian planets, and since the surface gravity of Saturn is almost precisely the same as for Neptune, the adiabatic gradient computed for this planet, 0·8° C. per km. (Kuiper 1949 b), should apply also here. The temperature increase within the 13·5 km. layer considered above will therefore be 11° C. This increase will in turn cause a steep gradient in the saturation pressure of NH_3; above the 13 km. layer the amount will be negligible, below it the amount will cause complete opacity to 9–13 μ even in thin layers. The radiometric observations should therefore give a value between 120° and 130° K., in agreement with experience (a numerical integration gives 127° K. for 1 cm. atm. of NH_3).

The argument will not hold for Uranus and Neptune; the top of the NH_3 gas will be so deep there that even extremely weak absorptions by CH_4, and possibly other gases in the 9–13 μ band, will succeed in forming an effective screen above the ammonia.

The visual penetration into the Saturn atmosphere is much deeper than the radiometric penetration. The latter is stopped by 1 cm. atm. of NH_3, while spectral observations around 7000 A. show about 200 cm. atm. of NH_3. The visual penetration is probably stopped by condensation products, possibly from ammonia. The case is not dissimilar to that presented by our own atmosphere; if we were to observe a cloud-covered Earth from the Moon in $\lambda = 6\,\mu$ (or in $20\,\mu < \lambda < 500\,\mu$) we would penetrate perhaps to the tropopause, while visually we would penetrate to the top of the overcast, many kilometres lower. We shall find a minimum depth and minimum temperature for the limit of visual penetration by assuming that NH_3 is saturated all the way down to the 'cloud' layer. Numerical integration shows this depth to be about 32 km. below the radiometric limit and to be at a temperature of about 153° K. We may estimate the pressure at this visual limit by making the plausible assumption that H_2 is 100 times as abundant as CH_4, i.e. that 35 km. atm. of H_2 lies above the limit (cf. Table 2). This will cause a pressure on Saturn of 0·35 atm. This pressure is probably of the right order of magnitude, while the actual temperature may be somewhat higher than $153° \text{ K.} = -120°$ C.

The above data show that if NH_3 is responsible for the 'cloud' layer the latter must be composed of ammonia crystals, not droplets, a conclusion already reached by Dunham (1934). We might speak of 'ammonia cirrus'. We would not expect the cirrus to have a sharp upper boundary, and there is no reason why Jupiter and Saturn should have precisely the same temperature at the limit of visual penetration and therefore show the same colour effects.

Each of the Jovian planets is expected to have a 'stratosphere', to which the Gold–Humphreys theory applies, perhaps with greater precision than to our

own stratosphere. The temperature T_s of this region is then given by $T_s^4 = \frac{1}{2} T_0^4$, if T_0 is the temperature of the effective radiating layer of the (lower) atmosphere. (As is well known, the above relation is due to the fact that the stratosphere receives its radiation from only one hemisphere; solar radiation is assumed to pass freely through the stratosphere.) The rôle of H_2O and CO_2 in the case of the Earth is played by CH_4 in the stratospheres of the Jovian planets, though the principal constituents must be H_2 and He. These upper layers are not expected to contain condensation products in view of the exponential decrease in density and constant temperature. The uppermost layers, however, will bear some resemblance to the terrestrial ionosphere and 'exosphere' (Spitzer 1949), and be hotter than the stratosphere below.

The atmospheric phenomena observed on Saturn appear from the Earth much less spectacular than those on Jupiter. This has usually been attributed to a lower temperature on Saturn, but the above discussion (pp. 268–269) should make us hesitate to accept this as an adequate explanation. The greater distance of Saturn will hide many details from view which would be visible at the distance of Jupiter; but this alone will not explain the full difference in appearance. Perhaps the most important single factor is the very rapid rotation of Saturn for its low mean density, and the great differential rotation between the different latitude zones (see below). As a result 'spots' will soon be stretched out into zones or belts.

Plate VIII shows three exquisite photographs of Saturn and its Rings, taken at the Lowell Observatory, and made available through the courtesy of Dr. E. C. Slipher. Prominent zones are shown in all three photographs, but important differences are noted. Reference is made to an early description of the zones and belts by Williams (1892). Of the 21 (bright) zones and (dark) belts seen by him at that time some zones showed a mottled appearance and some belts appeared reddish. Some rose colours and tints were reported by Holden (1891) who observed the planet with the Washington 26-inch refractor. During the 1950 opposition the writer, with the 82-inch telescope, found the S. equatorial belt light brown and the N. equatorial belt chlorine green. The S. polar cap was light brown but the N. polar cap was ivory.

The most prominent features on the planet are the occasional white spots, of which those of 1876, 1903 and 1933 appear to have been the brightest. Recently a similar spot has been reported by Danjon (1946). The 1933 spot has been observed both photographically, in different colours, by Wright (1933) and visually by several observers (Peek 1934). The 1933 spot was invisible on 1st August (E. C. Slipher 1934) and was discovered on 3rd August (by Hay and Weber), at which time its E–W diameter was one-fifth of the planet's equator. On 6th August its E–W diameter was one-quarter, on 18th August, two-thirds of the equator; on 24th August it had become a long white band, equal to half the planet's circumference. The period of rotation, initially 10h. 16m., decreased steadily till it was 10h. 13m. on 10th September. On 29th August, a second, small, spot appeared where the first one had formed on 3rd August. Wright found the spot invisible in ultra-violet towards the planet's limb, indicating a fairly deep location in the atmosphere. The spot was seen to merge into a previously existing bright zone, modifying the appearance of that zone appreciably. The zone was dark on Wright's ultra-violet photographs, which indicates an even lower level for it than for the top of the white spot. Wright interpreted the detailed structure of the spot and its gradual disintegration as an eruption

from a deep level; the eruptive centre was assumed to be marked by the ultra-violet photographs, which show the highest parts of the erupted material. A 'wind' at high levels, with a rotational velocity shorter than the spot, overtook the eruption and dispersed the products in a long tail emanating from the original eruption. Wright makes the interesting suggestion that this may be the mode of origin of the bright zones. A drawing of the 1933 spot has been published by E. C. Slipher (1938).

The 1876 and 1933 spots occurred near Saturn's equator and each gave a period of rotation of approximately 10h. 14m. The 1903 spot at 36° latitude gave a period of 10h. 38m.; another spot in the same belt, observed in 1910, gave about the same value. The large 'equatorial acceleration' indicated by these data was confirmed by a very beautiful spectroscopic investigation by Moore (1939). He determined the period of rotation at latitudes 0°, 27°, 42° and 57° (as measured at the centre of the spheroid) and found 1·06 times the equatorial period at 27° latitude; 1·08 and 1·11 at 42° and 57° latitude respectively. For the equatorial period of rotation he derived, adopting Struve's dimension for the planet's equator, 10h. 02m. ± 4m., 2% shorter than the value given by the spots. The ratios for the higher latitudes quoted above are independent of the adopted linear diameter, and are estimated to have uncertainties of the order of 1%. Moore's results pose a fascinating problem in hydrodynamics.

The orbital eccentricity is 0·056, which implies an equal percentage annual variation in mean temperature, or 5–6° C. Since our visual penetration may always be to a cirrus layer formed at a given temperature, there may be no observable effects of this temperature variation. The same may be true for the seasonal variation otherwise to be expected from the large tilt, 26·7°, of the equator on the plane of the orbit. To the writer's knowledge no seasonal phenomena have been recognized on Saturn.

The photographs by light of different colours (Wood 1916, Wright 1927 b) resemble those of Jupiter in general distribution. The reflectivity of the planet between H_β and H_α may be seen by comparison with the Ring from spectra by V. M. Slipher (1906); it is seen that the planet is yellowish, apart from the CH_4 absorptions.

A significant but puzzling result was the discovery by Becker (1933, 1948) of semi-regular, long-period fluctuations in the reflectivity of Saturn and the other Jovian planets. (The apparent fluctuations in the brightness of the planet due to its varying distances to the Earth and to the Sun were eliminated as well as the contribution of the Ring.) The average interval between consecutive minima was found to be of the order of 10 years; the observations comprise the interval 1860–1948, with gaps from 1888–1905 and from 1923–1934. The amplitude of the fluctuations is of the order of 0·4 mag. As Becker has pointed out, the bright spots of 1876, 1903 and 1933 all occurred during minima of reflectivity; the same is true of the 1946 spot. Becker found some slight correlation between minima in the reflectivity and minima in the solar activity (sunspot number), but it does not look as if solar activity is the only, or even the chief, cause of the phenomenon.

Uranus and Neptune

These planets are so distant that in general surface detail cannot be observed. Weak bands have been at times suspected on Uranus with the 36-inch Lick refractor (Holden *et al.* 1891) and have been seen at the Lowell Observatory. The most

important observations depending on the finite size of the disc are the spectroscopic determinations of the periods of rotation. They were made for Uranus by Deslandres (1902), Lowell and V. M. Slipher (1912) and Moore and Menzel (1930), and for Neptune by Moore and Menzel (1928). Reference is made to the pioneer work of V. M. Slipher at the Lowell Observatory in the observation of the red and infra-red spectra of the outer planets (1904, 1906, 1909, and 1933).

Reference has already been made to Becker's remarkable discovery of semi-regular, long-period fluctuations in the reflectivity of the Jovian planets. His data for Uranus have been confirmed by Ashbrook (1948) by the latter's independent observations. The average cycle is about 8·4 years and the amplitude is about 0·3 mag. As Ashbrook pointed out, it will be of interest to determine whether the equivalent path length of CH_4 undergoes corresponding changes. Stebbins and Jacobson (1928) made a careful photoelectric study of Uranus and found no short-period fluctuations, though they detected a variation of 0·01 mag. with phase (0° to 2·5°).

We have noted above the unidentified bands in the spectra of Uranus and Neptune (Kuiper 1949 c). The very large equivalent path of CH_4 observed in these planets may be attributed to the absence of condensation products in the upper layers. Some numerical data were given by Kuiper (1949 b).

Pluto

The magnitude and colour of Pluto have been determined by Baade (1934); the colour is slightly yellowish (colour index 0·67). No atmospheric absorptions have as yet been found, which is less surprising now that the small dimensions of this object are known. The ultra-violet spectrum, recently photographed by the writer, puts the upper limit on Pluto's atmosphere at about 0·1 terrestrial atmospheres.

General Remarks

The above Report is in no way a complete account of the literature on planetary atmospheres. The emphasis placed on certain problems will inevitably reflect the author's personal outlook. The reader will find an account of several of the topics treated here in the following references: Graff (1929), Russell (1935), Bobrovnikoff (1944) and Russell, Dugan and Stewart (1945). The problem of the origin of planetary atmospheres, not included here, is considered by Wildt (1939 b); the same author (1937) examines the photochemistry of planetary atmospheres. References to the Earth as a planet have been omitted, including such astronomical topics as the intensity distribution of refracted light on the Moon during a total lunar eclipse.

APPENDIX

Photographs of Venus by Messrs. Edson, Wright, Canwright and Winget, edited by A. B. Meinel

The photographs of Venus reproduced in Plates IX and X were obtained during the inferior conjunction of the planet in June 1940. Such photographs have to be made in day-time and require, therefore, a very clear sky. Table Mountain, California, was selected; the Smithsonian Institution has a Solar

observatory there and the elevation is 7,500 feet. The observations were made with a 6-inch reflector. Scattering in the instrument was minimized with the aid of a sun shield, supported on a 50-ft. cantilever tower built for the purpose. With careful guiding of the shield the telescope could be kept in the shade during the observations.

Venus was photographed on nine consecutive days, in seven wavelength bands ranging from 3100 to 9000 A. Only minor differences in the appearance of the cusps were noted between the different colours. Before minimum distance was reached two bright areas appeared in one of the cusps. These areas are seen on Plate X and on images 4, 6, 7 and 8 of Plate IX. The other cusp is extended on image no. 3. Plate X is a (reversed) composite print and corresponds to image 6 of Plate IX. It is probable that the intensity maxima and minima in the ring are caused by planet-wide circulation.

REFERENCES

ADAMS, W. S., and DUNHAM, T., Jr., 1932, *Publ. Astr. Soc. Pacif.*, **44**, 243, 380; 1937, *Ibid.*, **49**, 209.

ADEL, A., 1935, *Phys. Rev.*, **48**, 103; 1937, *Astrophys. J.*, **86**, 337; 1941, *Ibid.*, **93**, 397; 1949, *The Atmospheres of the Earth and Planets* (Chicago: University of Chicago Press), Ch. 10.

ADEL, A., and SLIPHER, V. M., 1934 a, *Nature, Lond.*, **134**, 148; 1934 b, *Phys. Rev.*, **46**, 240, 902; 1935, *Ibid.*, **47**, 651.

ANTONIADI, E. M., 1910, *J. Brit. Astr. Assoc.*, **20**, 189; 1919, *Ibid.*, **29**, 225; 1930, *La Planète Mars* (Paris: Hermann et Cie); 1933, *C. R. Acad. Sci., Paris*, **196**, 1087, *l'Astronomie*, **47**, 345; 1934, *La Planète Mercure* (Paris: Gauthier-Villars); 1935 a, *C. R. Acad. Sci., Paris*, **201**, 35; 1935 b, *J. Brit. Astr. Assoc.*, **45**, 236; 1939, *Sky*, **3**, No. 9, 6; 1941, *l'Astronomie*, **55**, 145, 241.

ASHBROOK, J., 1948, *Publ. Astr. Soc. Pacif.*, **60**, 116.

BAADE, W., 1934, *Publ. Astr. Soc. Pacif.*, **46**, 218.

BARNARD, E. E., 1895, *Pop. Astr.*, **2**, 433; 1911, *Mon. Not. R. Astr. Soc.*, **71**, 471, 714; 1914, *Astrophys. J.*, **40**, 259.

DE LA BAUME PLUVINEL, A., and BALDET, F., 1909, *C. R. Acad. Sci., Paris*, **149**, 838.

BECKER, W., 1933, *Sitzb. Preuss. Akad. Wiss. Phys.-Math. Kl.*, **28**, 839; 1948, *Astr. Nach.*, **277**, 65.

BOBROVNIKOFF, N., 1944, *Rev. Mod. Phys.*, **16**, 271–285.

BRITISH ASTRONOMICAL ASSOCIATION MEMOIRS, 1891–1943, *Reports of the Jupiter Section*, Vol. 1–35; *Reports on the Mars Section*, esp. 1916, **20**, 25–92, covering the 1990 opposition.

CAMICHEL, H., 1943, *l'Astronomie*, **57**, 49.

CAMPBELL, W. W., 1894, *Publ. Astr. Soc. Pacif.*, **6**, 139; 1918, *Ibid.*, **30**, 133.

CERULLI, V., 1898, *Bull. Astr., Paris*, **12**, 270.

COBLENTZ, W. W., 1925 a, *Astr. Nachr.*, **224**, 361; 1925 b, *Pop. Astr.*, **33**, 310, 363; 1925 c, *Nature, Lond.*, **116**, 439; 1927, *Ibid.*, **35**, 145; 1942, *J. Res. Nat. Bur. Stand., Wash.*, **28**, 297.

COBLENTZ, W. W., and LAMPLAND, C. O., 1923, *Lowell Bull.*, **85**, 91; 1924 a, *Pop. Astr.*, **32**, 570; 1924 b, *Publ. Astr. Soc. Pacif.*, **36**, 272; 1924 c, *Ibid.*, **36**, 274; 1925, *J. Franklin Inst.*, June and July; 1926, *Publ. Astr. Soc. Pacif.*, **38**, 355; 1927, *Sci. Papers, Bur. Stand.*, **22**, 237.

COBLENTZ, W. W., LAMPLAND, C. O., and MENZEL, D. H., 1925, *Pop. Astr.*, **33**, 297; 1926, *Astrophys. J.*, **63**, 177; 1927, *Publ. Astr. Soc. Pacif.*, **39**, 97.

DANJON, A., 1936, *Ann. Strasbourg*, **3**, Part 3; 1943, *l'Astronomie*, **57**, 161; 1946, *Ibid.*, **60**, 45; 1948, *C. R. Acad. Sci., Paris*, **227**, 652.

DESLANDRES, H., 1902, *C. R. Acad. Sci., Paris*, **135**, 472.

DOLLFUS, A., 1946, *C. R. Acad. Sci., Paris*, **60**, 132; 1947, *Ibid.*, **61**, 259; 1948 a, *C. R. Acad. Sci., Paris*, **226**, 996; 1948 b, *Ibid.*, **227**, 331, 383.

DUNHAM, T., Jr., 1949, *The Atmospheres of the Earth and Planets* (Chicago : University of Chicago Press), Ch. XI; 1933, *Publ. Astr. Soc. Pacif.*, **45,** 42, 202; 1934, *Ibid.*, **46,** 231; 1939, *Ibid.*, **51,** 253.

EVANS, J. E., and MAUNDER, E. W., 1903, *Mon. Not. R. Astr. Soc.*, **63,** 488.

FLAMMARION, C., 1909, *La Planète Mars*, Vol. II (Paris : Gauthier-Villars), pp. 316–18.

GENTILI, G., 1943, *l'Astronomie*, **57,** 49.

GERASIMOVIČ, B. P., 1937, *Poulkovo Bull.*, No. 127.

GRAFF, K., 1929, *Handbuch der Astrophysik*, **4,** Ch. 4, pp. 358–395.

HALE, G. E., 1910 a, *J. Brit. Astr. Assoc.*, **20,** 191; 1910 b, *Mon. Not. R. Astr. Soc.*, **70,** 175; 1911, *Ibid.*, **71,** 714.

HARGREAVES, F. J., 1928, *Mon. Not. R. Astr. Soc.*, **89,** 209; 1929, *Ibid.*, **89,** 703 (Addendum); 1945, *J. Brit. Astr. Assoc.*, **55,** 2.

HERBIG, G. H., 1945, *J. Brit. Astr. Assoc.*, **55,** 125.

HERZBERG, G., 1949, *The Atmospheres of the Earth and Planets* (Chicago : University of Chicago Press), Ch. XIII.

HESS, S. L., 1948, *Publ. Astr. Soc. Pacif.*, **60,** 289; 1950, *J. Met.*, **7,** No. 1.

HOFFMEISTER, C., 1940, *Astr. Nachr.*, **270,** 167.

HOLDEN, E. S., 1891, *Publ. Astr. Soc. Pacif.*, **3,** 16.

HOLDEN, E. S., SCHAEBERLE, J. M., and KEELER, J. E., 1891, *Publ. Astr. Soc. Pacif.*, **3,** 283.

HSI-YIN SHENG, BARKER, E. F., and DENNISON, D. M., 1941, *Phys. Rev.*, **60,** 786.

VAN DE HULST, H. C., 1949, *The Atmospheres of the Earth and Planets* (Chicago : University of Chicago Press), Ch. III.

JARRY-DESLOGES OBSERVATORY, 1908–1946, *Observations des Surfaces Planétaires*, Vols. 1–10.

JEFFERS, H. M., 1939, *Publ. Astr. Soc. Pacif.*, **51,** 292; 1945, *J. Brit. Astr. Assoc.*, **55,** 125.

KASTLER, A., 1949, *J. Chim. Phys.*, **46,** 72.

KUIPER, G. P., 1944, *Astrophys. J.*, **100,** 378; 1949 a, ed., *The Atmospheres of the Earth and Planets* (Chicago : University of Chicago Press); 1949 b, *Ibid.*, Ch. XII; 1949 c, *Astrophys. J.*, **109,** 540; 1949 d, *Ibid.*, **110,** 93.

LAMPLAND, C. O., 1924, *Pop. Astr.*, **32,** 570; 1941, *Astrophys. J.*, **93,** 401.

LOWELL, P., 1898, *Lowell Obs. Annals*, Vol. 1, 46; 1908, *Publ. Astr. Soc. Pacif.*, **20,** 100; 1909, *The Evolution of Worlds* (New York : Macmillan Co.); 1912, *Lowell Bull.*, No. 53, 17.

LOWELL OBSERVATORY, *Reports to the U.S. Weather Bureau*, 1948–50.

LYOT, B., 1929 a, *Annales, Paris-Meudon*, **8,** Part 1; 1929 b, *C. R. Acad. Sci., Paris*, **189,** 425; 1930, *Ibid.*, **191,** 703; 1937, *Ibid.*, **205,** 895; 1938, *l'Astronomie*, **51,** 45.

LYOT, B., *et al.*, 1943 a, *Astrophys. J.*, **97,** 75; 1943 b, *l'Astronomie*, **57,** 49; 1943 c, *Ibid.*, **57,** 67; 1945, *Astrophys. J.*, **101,** 254; 1946, *l'Astronomie*, **60,** 161.

MACDONALD, T. M., 1919, *J. Brit. Astr. Assoc.*, **29,** 225.

MENZEL, D. H., 1923, *Astrophys. J.*, **58,** 65; 1925, *Pop. Astr.*, **33,** 296; 1926, *Astrophys. J.*, **63,** 48; 1927, *Publ. Astr. Soc. Pacif.*, **39,** 26.

MENZEL, D. H., COBLENTZ, W. W., and LAMPLAND, C. O., 1925, *Pop. Astr.*, **33,** 297; 1926, *Astrophys. J.*, **63,** 177; 1929, *Publ. Astr. Soc. Pacif.*, **39,** 97.

MILLMAN, P., 1939, *Sky*, Vol. 3, No. 2, p. 10.

MINNAERT, M., 1946, *Bull. Astr. Inst. Netherlands*, **10,** 71.

MOORE, J. H., 1939, *Publ. Astr. Soc. Pacif.*, **51,** 274.

MOORE, J. H., and MENZEL, D. H., 1928, *Publ. Astr. Soc. Pacif.*, **40,** 234; 1930, *Ibid.*, **42,** 330.

PEASE, F. G., 1924, *Publ. Astr. Soc. Pacif.*, **36,** 345; 1925, *Ibid.*, **37,** 321.

PEEK, B. M., 1928, *Mon. Not. R. Astr. Soc.*, **89,** 209; 1934, *J. Brit. Astr. Assoc.*, **44,** 220; 1937, *Mon. Not. R. Astr. Soc.*, **97,** 574; 1939, *J. Brit. Astr. Assoc.*, **50,** 4; 1940, *Ibid.*, **51,** 6.

PETTIT, E., and NICHOLSON, S. B., 1923, *Publ. Astr. Soc. Pacif.*, **35,** 194; 1924 a, *Pop. Astr.*, **32,** 601; 1924 b, *Ibid.*, **32,** 614; 1924 c, *Publ. Astr. Soc. Pacif.*, **36,** 269; 1924 d, *Ibid.*, **36,** 227; 1925, *Pop. Astr.*, **33,** 299; 1930, *Astrophys J.*, **71,** 131; 1935, *Publ. Astr. Soc. Pacif.*, **47,** 215; 1936, *Astrophys. J.*, **83,** 84.

PHILLIPS, T. E. R., 1915, *J. Brit. Astr. Assoc.*, **26,** 10; 1918, *Ibid.*, **28,** 214; 1928, *Mon. Not. R. Astr. Soc.*, **89,** 209.

PICKERING, W. H., 1905, *Harvard Annals*, **53**, 155–71; 1914, *Pop. Astr.*, **22**, 409, Fig. 1; 1921 a, *Ibid.*, **29**, 1, Plates I and II; 1921 b, *Mars* (Boston); 1925 a, *Pop. Astr.*, **33**, 440; 1925 b, *Ibid.*, **33**, 579; 1925 c, *Publ. Astr. Soc. Pacif.*, **37**, 12; 1926 a, *Pop. Astr.*, **34**, 289, 296, Plates IX to XII; 1926, b *Ibid.*, **34**, 304, Plate XIII; 1926 c, *Ibid.*, **34**, 482.

QUÉNISSET, F., 1911, *C. R. Acad. Sci., Paris*, **153**, 1208; 1921, *Ibid.*, 1645, reprinted in 1922, *l'Astronomie*, **36**, 126; 1931, *Mon. Not. R. Astr. Soc.*, **91**, 947.

RABE, E.,1949, *Harvard Announcement Card*, No. 1051, 27 December, 1949.

RORDAME, A., 1922, *Pop. Astr.*, **30**, 137.

ROSS, F. E., 1926, *Astrophys. J.*, **64**, 243; 1928, *Ibid.*, **68**, 57.

ROUGIER, G., 1933, *Ann. Strasbourg*, **2**, Part 3; 1937, *Ibid.*, **3**, Part 5.

RUSSELL, H. N., 1899, *Astrophys. J.*, **9**, 284; 1916, *Ibid.*, **43**, 73, and references given there; 1935, *The Solar System and Its Origin* (New York: Macmillan Co.).

RUSSELL, H. N., and DANIEL, Z., 1907, *Astrophys. J.*, **26**, 69.

RUSSELL, H. N., DUGAN, R. S., and STEWART, J. Q., 1946, *Astronomy*, Vol. I, revised ed.

ST. JOHN, C. E., and NICHOLSON, S. B., 1922 a, *Pop. Astr.*, **30**, 229; 1922 b, *Astrophys. J.*, **56**, 380.

SCHOENBERG, E., 1940, *Astronomical Papers dedicated to E. Stroemgren* (Copenhagen: E. Munksgaard), p. 181.

SLIPHER, E. C., 1921 a, *Pop. Astr.*, **29**, 69; 1921 b, *Publ. Astr. Soc. Pacif.*, **33**, 127, 182; 1922, *Ibid.*, **34**, 215; 1923, *Pop. Astr.*, **31**, 86; 1924 a, *Ibid.*, **32**, 564; 1924 b, *Publ. Astr. Soc. Pacif.*, **36**, 255; 1927, *Ibid.*, **39**, 209; 1929 a, *Ibid.*, **41**, 256; 1929 b, *Pop. Astr.*, **37**, 326; 1931, *Publ. Astr. Soc. Pacif.*, **43**, 241; 1934, *Pop. Astr.*, **42**, 360; 1937, *Publ. Astr. Soc. Pacif.*, **49**, 140; 1938, *Proc. Amer. Phil. Soc.* **79**, 441; 1939, *Sky*, **3**, No. 5, p. 3; 1940 a, *Publ. Astr. Soc. Pacif.*, **52**, 284; 1940 b, *The Telescope*, **7**, 102.

SLIPHER, V. M., 1904, *Lowell Bull.*, No. 13, 87; 1906, *Ibid.*, No. 27, 173; 1909, *Ibid.*, No. 42, 231; 1912, *Ibid.*, No. 53, 19; 1924, *Publ. Astr. Soc. Pacif.*, **36** 261; 1927, *Ibid.*, **39**, 147; 1933, *Mon. Not. R. Astr. Soc.*, **93**, 657 (George Darwin Lecture).

SPITZER, L., 1949, *The Atmospheres of the Earth and Planets* (Chicago: University of Chicago Press), Ch. VII.

STEBBINS, J., and JACOBSON, T. S., 1928, *Bull. Lick Obs.*, **13**, 180.

TIKHOV, G. A., 1910, *Obs.*, **33**, 41; 1911, *Pulk. Mitt.*, **4**, No. 42; 1945, *C. R. Acad. Sci., l'URSS* (*Doklady*), **49**, No. 2; 1946, *Kasakh Ac. Sc.*, **4** (13) (Russian).

TRUMPLER, R. J., 1924, *Publ. Astr. Soc. Pacif.*, **36**, 263; 1927 a, *Ibid.*, **39**, 103; 1927 b, *Bull. Lick Obs.*, **13**, 19.

VAN BIESBROECK, G., 1924, *Pop. Astr.*, **32**, 589.

WESSELINK, A. J., 1948, *Bull. Astr. Inst. Netherlands*, **10**, 356.

WHIPPLE, F. L., 1941, *Earth, Moon and Planets* (Philadelphia: Blakiston Co.).

WILDT, R., 1931, *Naturwissenschaften*, **19**, 109; 1932, *Veröff. d. U. Sternw. Göttingen*, No. 22; 1937, *Astrophys. J.*, **86**, 321; 1939 a, *Mon. Not. R. Astr. Soc.*, **99**, 616; 1939 b, *Proc. Amer. Phil. Soc.*, **81**, 135; 1940 a, *Astrophys. J.*, **91**, 266; 1940 b, *Ibid.*, **92**, 247; 1942, *Ibid.*, **96**, 312.

WILLIAMS, A. S., 1892, *Mem. Brit. Astr. Assoc.*, **2**, 13; 1899, *Mon. Not. R. Ast. Soc.*, **59**, 376; 1920, *Ibid.*, **80**, 467; 1930, *Ibid.*, **90**, 696; 1936 a, *Ibid.*, **97**, 105; 1936 b, *J. Brit. Astr. Assoc.*, **47**, 68.

WOOD, R. W., 1916, *Astrophys. J.*, **43**, 310.

WRIGHT, W. H., 1924, *Publ. Astr. Soc. Pacif.*, **36**, 239; 1925, *Bull. Lick Obs.*, **12**, 48; 1926, *Publ. Astr. Soc. Pacif.*, **38**, 333; 1927 a, *Ibid.*, **39**, 220; 1927 b, *Ibid.*, **39**, 231; 1927 c, *Ibid.*, **39**, 358; 1927 d, *Bull. Lick Obs.*, **13**, 50; 1928, *Mon. Not. R. Astr Soc.*, **88**, 709 (George Darwin Lecture); 1929, *Ibid.*, **89**, 703; 1933, *Publ. Astr. Soc. Pacif.*, **45**, 236; 1939, *Ibid.*, **51**, 292.

WRIGHT, W. H., and KUIPER, G. P., 1935, *Publ. Astr. Soc. Pacif.*, **47**, 92.

PARAMAGNETIC RELAXATION EFFECTS

By A. H. COOKE
The Clarendon Laboratory, Oxford

CONTENTS

PAGE
§ 1. Introduction 276
§ 2. Theory of paramagnetic relaxation 278
§ 3. Spin–lattice relaxation 279
Interpretation of spin–lattice relaxation times 281
§ 4. Spin–spin relaxation 283
Interpretation of spin–spin relaxation times 284
§ 5. Experimental methods 287
Absorption measurements at radio frequencies 287
Dispersion measurements 288
Measurements at liquid helium temperatures 289
§ 6. Specific heats 291
Interpretation of magnetic specific heats 292
References 293

§ 1. INTRODUCTION

In the theory of paramagnetism, as it has developed since the classic work of Langevin, a paramagnetic substance is regarded as an assembly of magnetic dipoles, which in the absence of a magnetic field are distributed in directions at random, and so show no external magnetic moment. When a magnetic field is applied, the resulting magnetization is determined by the state of equilibrium between the magnetic forces tending to orient the dipoles in the direction of the field, the random forces of thermal agitation, and the constraints arising from the interactions of the dipoles with their surroundings and with one another. Since the theory is concerned only with the equilibrium states, no consideration is given to the time required for the establishment of equilibrium if, for example, the magnetic field is abruptly changed. Experimentally it is found that in this case the approach to the new equilibrium can be described in terms of one or more characteristic relaxation times, defined by equations of the type

$$\frac{dM}{dt} = \frac{M_0 - M}{\tau}. \qquad \ldots\ldots(1)$$

Here M_0 is the equilibrium value of the magnetic moment in a given direction, M is the actual value, and τ is a constant—the relaxation time. Although these time effects are seldom of practical importance, the relaxation times being generally a very small fraction of a second, it is worth noting that very long times are possible. Consider an assembly of free magnetic dipoles, distributed at random without any interaction with their surroundings or with one another. The application of a magnetic field will cause each dipole to precess about the direction of the field without any change in the component of its magnetic moment in this direction. The magnetic moment of the assembly in the direction of the field therefore remains zero. Only if some mechanism exists by which a dipole can change its energy of precession and turn towards the direction of the field can a net magnetic moment arise.

The substances which most nearly conform to this picture of free dipoles are the nuclear paramagnetics. For example, if a specimen of ice at liquid air temperature is placed in a magnetic field, a time of several hours is required for the development of a magnetic moment by the orientation of the protons of its water molecules (Turner, Sachs and Purcell 1949). Ice is not usually thought of as possessing a paramagnetic susceptibility at all (its nuclear paramagnetic susceptibility is only 3×10^{-9}) but it is important to note that the long equilibrium time arises not simply because the paramagnetism is feeble, but because of the extremely weak contact between the nuclear magnetic moments of the protons in ice and their surroundings. If by some means or other this contact is improved the equilibrium time will decrease. For example, if the ice is melted to water the relaxation time falls to a few seconds.

Of the substances usually thought of as paramagnetics, the most important are the paramagnetic crystals such as the salts of the elements of the iron group. These are characterized by containing an ion having an incomplete inner shell, with which is associated a magnetic dipole moment. Some of these ions such as Fe^{+++} or Mn^{++} possess only a spin magnetic moment, others such as Cr^{+++} or Cu^{++} possess both a spin moment and an orbital moment in the free state. In the crystal, however the orbital magnetism is 'quenched', the interaction of the orbital magnetic moment with the electric field of the crystalline lattice being so strong that no transitions between different orbital states are possible at ordinary temperatures. The paramagnetism of iron group salts can therefore be considered to be due to the spin moments, which are relatively weakly coupled to their surroundings, so that they approximate to an assembly of free dipoles. For instance, the susceptibility is inversely proportional to the temperature, which is the law of variation for an assembly of free dipoles. Their contact with their surroundings is, however, much stronger than that of the nuclear dipoles, and the relaxation times range from 10^{-10} to 10^{-6} second at room temperature.

This report will be concerned with relaxation effects in paramagnetic crystals. The first work to be published on this subject was Waller's theoretical study (1932). Waller showed that the phenomena fall into two classes, according as the applied magnetic field is weak or strong compared with the interaction field of the crystal's magnetic ions. In the first case the relaxation effects depend only on spin–spin interaction, and the relaxation time is of the order 10^{-10} second independent of the temperature. Waller outlined a satisfactory theory of this effect, which has since been generalized by Broer (1943). In the second case magnetization can occur only by the interaction of the spins with the crystal lattice. The relaxation times are much longer, and are temperature dependent, increasing as the temperature falls. Waller indeed deduced relaxation times much longer than the true values, because he neglected an important mechanism of interaction between the spins and the lattice, but his paper established the essential distinction between the two classes of phenomena, and the conditions under which each will occur. Further theoretical investigations of spin–lattice relaxation by Kronig (1939) and by Van Vleck (1940) have given relaxation times of the order of magnitude observed experimentally, though some phenomena remain unexplained.

Besides giving information on time effects, relaxation measurements afford a method of determining the magnetic specific heat of a paramagnetic, i.e. that part of the specific heat arising from the internal energy of the spin system. Because of interactions the different energy levels of the spins are slightly separated.

As the temperature is raised, the population of the higher energy levels rises, resulting in a magnetic specific heat. The interest in determining this specific heat lies in the information it gives on the spin energy levels. It is, however, extremely small, and for most paramagnetic salts becomes an appreciable fraction of the total specific heat only at temperatures near 1° K. Spin-lattice relaxation measurements, however, effectively isolate the magnetic specific heat, and so enable it to be determined at temperatures where direct calorimetric determinations would be out of the question. It was the thermodynamic theory of Casimir and Du Pré (1938) which first revealed the possibilities of this method of measurement, which has since been applied to very many paramagnetic salts.

The pioneer of experimental work was Gorter, whose researches, begun in Haarlem in 1935 and 1936, were the first to show the existence of spin–spin relaxation (Gorter 1936a) and spin–lattice relaxation (Gorter and Brons 1937). His monograph (Gorter 1947) is the standard work on the subject. Under his inspiration paramagnetic relaxation has been investigated in Holland at Haarlem, Groningen, Amsterdam, and especially at the Kamerlingh Onnes Laboratory at Leiden. Measurements have also been made in the United States of America and in England.

§2. THEORY OF PARAMAGNETIC RELAXATION

Paramagnetic relaxation measurements are always made by studying the behaviour of the salt in an oscillatory field, a large parallel steady field being superposed on the oscillatory field when spin–lattice effects are to be examined. As the frequency of the oscillatory field is raised, relaxation effects may be detected by the appearance of a component of magnetization out of phase with the field, accompanied by a diminution of the in-phase component. If we put $dM/dH = \chi$, the differential susceptibility of the salt, the existence of in-phase and out-of-phase components of magnetization can be represented by writing $\chi = \chi' - i\chi''$. The experimental results can then be expressed by the variation of χ' and χ'' with the frequency of the applied field. The first task of the theory is to relate the values of χ' and χ'' to the relaxation times, the second to relate the relaxation times to the atomic and lattice properties of the crystal.

The problem is simplified by the fact that spin–spin and spin–lattice relaxation effects can usually be studied separately. In all the salts which have been investigated the spin–spin relaxation times are of the order 10^{-10} second and so have very little effect on the behaviour of the salts at the frequencies of 10^7 c/s. or less which are used in the investigation of spin–lattice relaxation. On the other hand, spin–lattice effects do not occur in small fields, in which magnetization occurs by polarization of the spin energy levels instead of by transfer of spins from higher to lower energy levels. In any paramagnetic, each spin is subject to a fluctuating magnetic field, of the order μ/d^3 where μ is the magnetic moment of each spin and d is the distance separating neighbouring spins. In iron ammonium alum, for instance the r.m.s. value of this internal field is 450 oersteds (Volger, de Vrijer and Gorter 1947a). If an external field very much smaller than the internal field is applied, its chief effect is to change the direction but not the magnitude of the field on each dipole. Magnetization then occurs by a slight re-orientation of the dipoles, without any interchange of energy between them and the lattice. The energy of magnetization is small compared with the energy of interaction of the dipoles and is taken up by small adjustments of their energy

levels. If on the other hand a magnetic field is applied which is large compared with the internal field, a small increase of the external field changes the magnitude but not the direction of the field on each dipole. The resulting increase in the magnetic moment of the specimen occurs by the transfer of dipoles from positions in which their moments oppose the field to positions in which they assist it, the energy difference between these positions being given to the lattice. At intermediate values of magnetic field both processes contribute to the magnetization, but clearly only the second process (change of population of the energy levels) involves spin–lattice relaxation.* The two relaxation effects can therefore be studied separately, spin–spin effects by measurements at high frequencies in small magnetic fields, spin–lattice effects by measurements in larger magnetic fields, at frequencies so low that spin–spin relaxation does not affect the results.

§3. SPIN–LATTICE RELAXATION

Spin–lattice relaxation effects can be treated thermodynamically. The basis of the theory is that the relaxation time for the establishment of equilibrium among the spins is so short that the spin system is always in thermodynamic equilibrium and so can be treated as a thermodynamic entity separate from the crystal lattice, possessing for example, its own specific heat and a temperature of its own which may differ from that of the lattice. If now a large steady field and a parallel oscillatory field are applied to a paramagnetic salt, the heat of magnetization will cause the temperature of the spin system to oscillate with the field, rising as it increases and falling as it decreases. These oscillations of temperature will be reduced by the thermal contact between the spins and the lattice; if the period of the oscillatory field is sufficiently long compared with the spin–lattice relaxation time, the temperature of the spins will never differ appreciably from that of the lattice (supposed constant) and the susceptibility of the salt will have its isothermal value χ_T. On the other hand, if the period is too short for any appreciable heat transfer between spins and lattice, the susceptibility will have a different value, which we will call the adiabatic value, χ_S.

$$\chi_T=\left(\frac{\partial M}{\partial H}\right)_T; \qquad \chi_S=\left(\frac{\partial M}{\partial H}\right)_S. \qquad \text{.....(2)}$$

Two thermodynamic relations follow without any further assumptions about relaxation effects:

$$\chi_T/\chi_S=C_H/C_M, \qquad \text{.....(3)}$$

and

$$C_H-C_M=T\left(\frac{\partial M}{\partial T}\right)_H^2\frac{1}{\chi_T}. \qquad \text{.....(4)}$$

Here M is the magnetization in a field H, and C_H and C_M are the specific heats of the spin system at constant field and at constant magnetization

* The distinction between the two processes is clear in the following simple example. Consider a system of N dipoles each of magnetic moment μ, which can lie either parallel or anti-parallel to a magnetic field. The usual statistical calculation shows that in a field H, such that $\mu H \ll kT$, the excess in the parallel position is $N\mu H/kT$, and the net magnetic moment is $N\mu^2H/kT$. If now a small field h is applied perpendicular to H the dipoles will re-orient themselves along the resultant of h and H, and a few will pass from the anti-parallel to the parallel position to give a net magnetic moment in this direction $M'=N\mu^2(H^2+h^2)^{1/2}/kT$. Of this magnetic moment the component in the direction of h is $N\mu^2h/kT$. Thus the susceptibility in the direction of h is $N\mu^2/kT$, as before, but it arises partly from transitions from the anti-parallel states to the parallel states, and partly by a polarization of the states in the direction of h. If $h \ll H$, the magnetization arises entirely from the polarization effect, the number of dipoles which change their energy level being proportional to $[(H^2+h^2)^{1/2}-H]$, which vanishes to a higher order than h.

respectively. These equations are the basis of the method of employing relaxation effects for the determination of specific heats from magnetic data.

Values of χ_T and χ_S of course give no information on relaxation effects. To describe these, or to determine specific heats under conditions in which χ_S cannot be measured because the frequency required is too high, further information on the heat transfer between spins and lattice is needed. Casimir and Du Pré (1938) assumed that the rate of flow of heat at any instant is directly proportional to the temperature difference between the spin system and the lattice. Their theory has been generalized by Debye (1938) whose treatment is followed here.

When the spin temperature exceeds the lattice temperature by ΔT, the rate of flow of heat to the spin system is

$$\frac{dQ}{dt} = -\alpha \Delta T. \qquad \ldots\ldots(5)$$

Integrating, $$-\alpha \int \Delta T\, dt = \Delta Q = C_M \Delta T - T\left(\frac{\partial H}{\partial T}\right)_M \Delta M.$$

If the field varies periodically with a frequency $\omega/2\pi$, the temperature variation will have the same frequency, and

$$-\frac{\alpha}{i\omega}\Delta T = C_M \Delta T - T\left(\frac{\partial H}{\partial T}\right)_M \Delta M.$$

But $$\Delta M = \left(\frac{\partial M}{\partial H}\right)_T \Delta H + \left(\frac{\partial M}{\partial T}\right)_H \Delta T.$$

Hence, eliminating ΔT,

$$\frac{\Delta M}{\Delta H} = \left(\frac{\partial M}{\partial H}\right)_T \frac{C_M + \alpha/i\omega}{C_M - T\left(\frac{\partial H}{\partial T}\right)_M \left(\frac{\partial M}{\partial T}\right)_H + \frac{\alpha}{i\omega}}$$

or $$\chi = \chi_T \frac{C_M + \alpha/i\omega}{C_H + \alpha/i\omega}. \qquad \ldots\ldots(6)$$

Separating the real and imaginary parts of $\chi = \chi' - i\chi''$,

$$\chi' = \frac{\chi_T - \chi_S}{1 + \omega^2\tau^2} + \chi_S, \qquad \ldots\ldots(7)$$

$$\chi'' = (\chi_T - \chi_S)\frac{\omega\tau}{1 + \omega^2\tau^2}, \qquad \ldots\ldots(8)$$

where $$\tau = C_H/\alpha. \qquad \ldots\ldots(9)$$

Provided that the assumption that the heat flow is proportional to the temperature difference between the spins and the lattice is valid, these formulae enable the spin–lattice relaxation time τ and the adiabatic susceptibility χ_S to be determined either from dispersion measurements, which give χ', or from measurements of the heat absorbed in an oscillatory field, which give χ''. The heat absorbed per second per gramme of salt in an oscillatory field of amplitude H_1 is

$$A = \tfrac{1}{2}\omega\chi'' H_1^2.$$

Figure 1 shows the variation with frequency of χ' and χ'', for the case $\chi_S = 0{\cdot}2\chi_T$, and also the absorption A, in arbitrary units.

Broadly speaking, the Casimir–Du Pré theory has been verified in a great many experiments, and has been applied successfully to the measurements of relaxation

times and specific heats. Recent experiments at liquid helium temperatures (Kramers, Bijl and Gorter 1950, Benzie and Cooke 1950a) have suggested, however, that the relaxation effects at very low temperatures cannot be expressed in terms of a single relaxation time.

Interpretation of Spin–Lattice Relaxation Times

The spin–lattice relaxation time τ can be related to the probability of transitions between different energy levels of the spin system being induced by interaction with the lattice. Since the more important transitions are those which involve large energy changes, the relaxation time is the inverse of a suitably weighted mean of the different transition probabilities.

$$\frac{1}{\tau} = \frac{\Sigma A_{12} W_{12}^{\,2}}{\Sigma W_{12}^{\,2}}. \qquad \ldots\ldots(10)$$

Here A_{12} = transition probability from state 1 to state 2 due to interaction with the lattice. W_{12} = energy difference of states 1 and 2. This relation, which

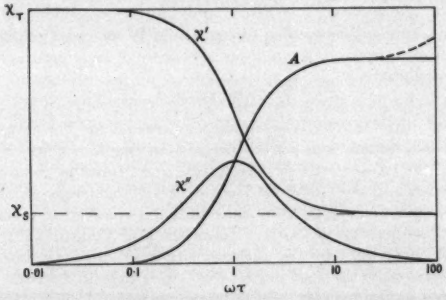


Figure 1. Effect of spin–lattice relaxation on the susceptibility of a paramagnetic salt at different frequencies. χ' = susceptibility in-phase with the oscillatory field. χ'' = susceptibility out-of-phase with the oscillatory field. A = heat absorption per second, in arbitrary units. The dotted curve shows the additional heat absorption at high frequencies due to spin–spin relaxation.

was used by Van Vleck (1940), may be derived formally (Broer 1945, Gorter 1947) from the theory of Gorter and Kronig (1936).

Interaction between the spin system and the lattice arises from the distortion of the lattice by thermal vibration, which may affect the spin system in a number of ways. Waller (1932) in his original paper pointed out that two types of process must be considered, the direct absorption or emission of a quantum of lattice vibrations by the spin system, and the inelastic scattering of the lattice vibrations by the absorption of one quantum and the emission of another of different energy. This second process, which may be compared with the Raman effect in optics, is especially important at high temperatures, since it allows the spin system to interact with vibrational quanta of much higher energy than those which take part in the direct process. The only mechanism of interaction which was considered by Waller was the effect of the lattice vibrations in modulating the magnetic interaction field of the spins by varying their distance apart. This interaction is very

small and its contribution to the transition probabilities is, in fact, negligible. Heitler and Teller (1936) and Fierz (1938) remarked that since the spin energy levels are influenced by the Stark effect of the electric field of the crystalline lattice, the variations of the electric field due to thermal vibration provide a perturbation capable of inducing transitions. Their calculations showed, however, that this mechanism is still inadequate to account for the observed relaxation times.

Kronig (1939) showed that the electric field of the lattice exerts a more powerful influence through the spin–orbit coupling. When the lattice vibrates, the orbital motion in the magnetic ions undergoes periodic changes due to the variations in the electric field of the crystal, and these changes react on the spins through the spin–orbit coupling, and may enable them to alter their orientation with respect to the constant external magnetic field. A distinguishing feature of this interaction is its strong dependence on the separation of the excited states of the orbital motion of the magnetic ions from the ground state. For the simplest case of ions with $s=\frac{1}{2}$ (for example Ti^{+++}), Kronig obtained for the direct absorption process a transition probability

$$B \propto A^2H^4T/W^4,$$

where A is the spin–orbit coupling constant, and W the energy separation of the lowest excited state from the ground state. For the Raman process he obtained a transition probability

$$D \propto A^2H^2T^7/W^6,$$

provided $T \ll \theta$, the Debye characteristic temperature for the lattice vibrations. Assuming $W=1{,}000\,cm^{-1}$ for Ti^{+++} in caesium titanium alum, Kronig found $B \simeq 10^{-10}H^4T$, and $D \simeq 10^{-9}H^2T^7$, giving relaxation times of about 10^{-11} second at 1,000 oersteds and 90° K. and 10^{-2} second at 1,000 oersteds and 1° K. These figures would suffice to explain the failure to detect spin–lattice relaxation effects in caesium titanium alum. Gorter, Teunissen and Dijkstra (1938) gave 10^{-9} second as the upper limit for the relaxation time in their experiments at liquid air temperatures, while de Haas and Du Pré (1938) gave 10^{-2} second as an upper limit from their experiments at helium temperatures.

Van Vleck (1940) showed, however, that when the problem is considered in greater detail the agreement with experiment is less satisfactory. He made explicit calculations for caesium titanium alum and potassium chrome alum of the interaction of a magnetic ion with the vibrations of the surrounding complex of six water molecules. The relaxation times obtained differ considerably from the 'order of magnitude' figures of Kronig, not only in their numerical values but also in their dependence on magnetic field. For caesium titanium alum, taking the separation of the orbital levels as $1{,}000\,cm^{-1}$, the direct absorption processes are found to be the more important at helium temperatures, giving relaxation times at 1·2° K. of 5×10^4 second in zero field, $1{\cdot}7 \times 10^2$ second at 1,000 oersteds and $1{\cdot}8 \times 10^{-2}$ second at 10,000 oersteds. The agreement with experiment, Van Vleck remarks, is miserable. At liquid air temperatures the Raman processes preponderate and give relaxation times of 10^{-9} second at 77° K. Recent experimental work has shown that this time is also too long. Measurements by the paramagnetic resonance method (Bagguley *et al.* 1948) indicate that the relaxation time is less than 10^{-11} second at temperatures down to 20° K., while measurements by Bijl (1950) have shown that at 6 to 7° K. the relaxation time is 10^{-10} second. To account for these results by Van Vleck's theory it would be

necessary to take a value of 100 cm^{-1} instead of 1,000 cm^{-1} for the orbital energy difference, W. The calculated relaxation times would then be 2×10^{-4} second at 1·2° K., 3×10^{-11} second at 7° K., and at 77° K. would be much smaller than the spin–spin relaxation time. Since Bijl's measurements give $W = 400$ cm^{-1} (Bleaney 1950) there remains some discrepancy between theory and experiment.

Van Vleck's calculated relaxation times for chromium potassium alum are in fair agreement with the experimental results in order of magnitude, but do not correctly predict the dependence on magnetic field and temperature at liquid helium temperatures. For example, at 1·4° K. Van Vleck obtains relaxation times of 0·011 second in zero magnetic field, falling as the magnetic field is increased to 0·0030 second at 3,000 oersteds. At a given field, the relaxation time should be inversely proportional to the temperature. It is found experimentally that different specimens give different relaxation times, of the order of 0·01 second, but that the results are consistent in that the relaxation times increase slowly with increasing magnetic field, and vary with temperature, at a rate between $T^{-3/2}$ and T^{-2}. At liquid air temperatures, on the other hand, the agreement of theory and experiment is very good. At 90° K. the observed relaxation times are $0{\cdot}06 \times 10^{-6}$ second at 800 oersteds increasing to $0{\cdot}19 \times 10^{-6}$ second at 3,200 oersteds against calculated values of $0{\cdot}11 \times 10^{-6}$ second and $0{\cdot}19 \times 10^{-6}$ second respectively.

To summarize, when allowance is made for the extreme complexity of the calculations, it appears that the mechanisms suggested for spin–lattice relaxation are adequate to account for the relaxation times observed at liquid air temperatures, but that at liquid helium temperatures several features of the phenomena are unexplained, in particular the existence of different relaxation times for the same substance, and the fact that the relaxation times are not inversely proportional to temperature. The difficulties have been discussed by Van Vleck (1940). Temperley (1939, 1947) has suggested that because of spin–spin interaction the spins may undergo transitions in groups rather than singly as is supposed in the theory. Experiments on diluted salts show that the relaxation times at helium temperature are longer than for the pure salts (Bijl 1941) but quantitative evidence is lacking.

de Vrijer and Gorter (1948) have reported a new relaxation phenomenon in crystals of potassium chrome alum in fields of a few hundred oersteds at helium and hydrogen temperatures. At these temperatures the spin–lattice relaxation time is of the order of 10^{-4} second. The new effect is characterized by a relaxation time of the order of 10^{-7} second independent of temperature. It is anisotropic, depending on the angle between the magnetic field and the crystal axes. They suggest tentatively that it may arise from the finite time for the establishment of equilibrium between different groups of energy levels, for example, groups of large energy differences and small energy differences.

§ 4. SPIN–SPIN RELAXATION

When a paramagnetic salt is subjected to an oscillatory field of sufficiently high frequency, it is found that an absorption of heat additional to the spin–lattice absorption can be detected, which can be described by the equation

$$A_S = \tfrac{1}{2}\omega^2 \tau_S \chi_S H_1^2. \qquad \ldots\ldots(11)$$

Here A_S is the absorption of energy per gramme per second, H_1 the amplitude of the oscillatory magnetic field of frequency $\omega/2\pi$ and χ_S the adiabatic susceptibility of the salt per gramme. Since the absorption is proportional to χ_S, that is

to the susceptibility remaining when spin–lattice energy transfer is excluded, it is ascribed to spin–spin relaxation effects. τ_S has the dimensions of time, and is called the spin–spin relaxation time. It is found to be of the order of magnitude 10^{-10} second, is independent of temperature, and is little affected by the presence of a steady field parallel to the oscillatory field. With a perpendicular steady field, τ_S decreases as the steady field is increased.

Because τ_S is so small, spin–spin effects can be detected only when measurements are made at very high radio frequencies. No quantitative measurements of the departure of the absorption from a square law dependence on frequency have been made, nor have dispersion measurements been reported. The determination of τ_S therefore depends on measurements of the absolute value of the absorption, in contrast with spin–lattice measurements where relative measurements suffice.

Interpretation of Spin–Spin Relaxation Times

The notion of a spin–spin relaxation time is less obvious than that of a spin–lattice relaxation time. It first appears as a convenient form for the constant in the absorption equation (11), chosen by analogy with the limiting form at low frequencies of the Casimir–Du Pré equation for spin–lattice absorption. It is interpreted as a time constant characterizing the establishment of thermal equilibrium among the spins when this is disturbed by the application of a magnetic field. Two mechanisms of interaction between the spins may be expected to contribute to this, magnetic dipole interaction and exchange interaction. Some idea of their relative importance can be gained from elementary considerations. The effect of magnetic interaction is that each spin is subject to a varying magnetic field due to its neighbours. The r.m.s. value of this field is readily calculated (Van Vleck 1937) and is usually called the internal field, H_i, characteristic of the salt. Very approximately, it may be expected that the relaxation time due to magnetic interaction will be $\tau_S \simeq 1/\omega_L$, where $\omega_L/2\pi$ is the frequency of the Larmor precession of the spins in this field. If we write $\tau_S \simeq \gamma H_i$ where γ is the magnetogyric ratio ($\boldsymbol{e/mc}$ for free spins), we find that for iron alum, having $H_i = 450$ oersteds, $\tau_S \simeq 1{\cdot}2 \times 10^{-10}$ second, against the measured value $1{\cdot}1 \times 10^{-10}$ second (Volger 1946) and for gadolinium sulphate, having $H_i = 1380$ oersteds, $\tau_S \simeq 4{\cdot}1 \times 10^{-11}$ second, against a measured $5{\cdot}7 \times 10^{-11}$ second. On the other hand, in copper ammonium sulphate $H_i = 200$ oersteds, giving $\tau_S \simeq 2{\cdot}7 \times 10^{-10}$ second against a measured value of $9{\cdot}1 \times 10^{-10}$ second, and in copper sulphate $H_i = 370$ oersteds, giving $\tau_S \simeq 1{\cdot}5 \times 10^{-10}$ second against a measured value of $6{\cdot}7 \times 10^{-10}$ second. Exchange interaction is known to be important in these two salts (Pryce 1948, Benzie and Cooke 1949). Its effect is seen to be to lengthen the relaxation time, i.e. to increase the absorption.

This effect of exchange interaction has recently been considered theoretically by Wright (1949) following similar work by Van Vleck (1948). The spin–spin relaxation due to magnetic interaction was first calculated by Waller (1932), who considered the case of a salt in which there is no splitting of the magnetic energy levels of the spins by the electric field of the crystalline lattice (for example Ti^{+++} and Cu^{++} salts). Broer's treatment, which is followed here, took into account, at least approximately, the effect of electric splittings as well as the effect of a steady magnetic field superposed on the oscillatory field. We consider first the simple case in which these effects are absent.

Magnetic interaction will cause the magnetic ions to be distributed in energy over a range approximately equal to μH_1 where μ is the magnetic moment per ion. Transitions between two states of a magnetic ion, of energies W_p and W_q, may be induced by an oscillatory field of frequency ν, if $h\nu = |W_p - W_q|$, producing absorption or stimulated emission. The net absorption of energy will be proportional to the square of the oscillatory field, to the difference of occupation of the levels p and q, and to the square of the matrix element of magnetic moment in the direction of the field M_{pq}. If $|W_p - W_q| \ll kT$, the difference of population of the levels is $h\nu/2kT$. The absorption of energy per gramme per second can be shown to be

$$A_S = \frac{\omega^2 H_1^2}{8kT} \Sigma |M_{pq}|^2 = \frac{\omega^2 H_1^2}{8kT} f(\nu). \qquad \ldots\ldots(12)$$

The summation is taken over all pairs of levels whose energy difference lies in the range $h\nu$ to $h(\nu + \Delta\nu)$. Then comparing equations (11) and (12) we see that

$$\tau_S = \frac{f(\nu)}{4\chi_0 kT}. \qquad \ldots\ldots(13)$$

(In the absence of a steady magnetic field, $\chi_S = \chi_T = \chi_0$, the static susceptibility.) The problem is then to determine the distribution function $f(\nu)$. Broer shows that it must satisfy the conditions

$$\int_0^\infty f(\nu)\, d\nu = kT\chi_0 \qquad \ldots\ldots(14)$$

and

$$\int_0^\infty \nu^2 f(\nu)\, d\nu = kT\chi_0 \nu_0^2 \qquad \ldots\ldots(15)$$

where

$$h\nu_0 = g\beta H_1 \qquad \ldots\ldots(16)$$

where g is the Landé factor and β is the Bohr magneton.

These would determine $f(\nu)$ if its shape were known. Broer chose a Gaussian form, by analogy from Heisenberg's theory of ferromagnetism. Applying the above conditions we have

$$f(\nu) = \frac{kT\chi_0}{\nu_0} \sqrt{\left(\frac{2}{\pi}\right)} \exp\left(-\frac{\nu^2}{\nu_0^2}\right). \qquad \ldots\ldots(17)$$

The frequencies at which relaxation measurements are made are low compared with ν_0 (e.g. for iron ammonium alum $\nu_0 = 1{\cdot}3 \times 10^9$ c/s.) and we can therefore write

$$\tau_S = \frac{f(0)}{4\chi_0 kT} = \frac{1}{\sqrt{(8\pi)}} \frac{1}{\nu_0} = \frac{1}{\sqrt{(8\pi)}} \frac{h}{g\beta H_1}. \qquad \ldots\ldots(18)$$

Volger, de Vrijer and Gorter (1947 a) measured the spin–spin relaxation times of a number of paramagnetic salts, and have given a comparison with Broer's theory. In most cases fairly close agreement is obtained, though the experimental values are always larger than the theoretical ones. The discrepancy is greatest for salts in which exchange coupling is important.

Broer's theory shows that when spin–spin relaxation takes place in a steady magnetic field H_p perpendicular to the oscillatory field, the effect is to alter the form of the distribution function, shifting its maximum from zero frequency to a frequency such that $h\nu = g\beta H_p$. The absorption at low frequencies is therefore

decreased and, if magnetic interaction alone is considered, should fall to half its zero field value when the perpendicular field is approximately equal to H_l. Comparison with Volger's experiments shows that this is approximately true, again with the exception of the salts in which exchange is important. In these the field required to halve the absorption is much smaller than H_l.

The influence of exchange interaction has been discussed by Wright (1949) who has shown that in its absence the Gaussian form adopted for $f(\nu)$ by Broer is a good approximation, but that when exchange interaction is present the effect is to sharpen the maximum of $f(\nu)$, raising its peak above the value calculated by Broer, and narrowing its frequency range. If no perpendicular magnetic field is applied, so that the maximum of $f(\nu)$ lies at zero frequency, the absorption is increased by exchange interaction, but a comparatively small perpendicular field suffices to shift the maximum of $f(\nu)$ sufficiently to reduce the absorption to one half.

For the simplest case of exchange between a spin and six equidistant neighbours (cubic symmetry) Wright has calculated the effect of exchange on the spin–spin relaxation time, in terms of the exchange constant A defined by the equation

$$V_{ij} = A\boldsymbol{\mu}_i . \boldsymbol{\mu}_j / d^3. \qquad \ldots\ldots (19)$$

Here V_{ij} = energy of exchange interaction between two neighbouring spins i and j, of magnetic moments $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_j$, separated by a distance d. A is an approximate measure of the ratio of the exchange interaction energy to the magnetic interaction. The result is to increase the relaxation time from τ_S, the value in the absence of exchange, to τ_S' where

$$\tau_S' = \tau_S(1 + 0{\cdot}054A^2). \qquad \ldots\ldots (20)$$

Using this equation, Wright calculated the values of A required to account for the discrepancy between Volger's experimental results and Broer's calculated values for the spin–spin relaxation times in zero field. The values so derived may be compared with the values calculated from specific heat and susceptibility data (Van Vleck 1937, Opechowski 1937). All the calculations involve approximations (in particular the assumption of cubic symmetry), but agreement as to order of magnitude is obtained. Some typical results, both for substances in which exchange is believed to be important and for others in which it is not, are given in Table 1, in which the experimental relaxation times are compared with the values from Broer's theory.

Table 1. Data on Spin–Spin Relaxation Times

Substance	τ_S (expt.)	τ_S (theory)	H_l	H_p	A_1	A_2
$Fe(NH_4)(SO_4)_2 . 12H_2O$	1·1	0·8	450	870		
$Mn(NH_4)_2(SO_4)_2 . 6H_2O$	1·3	0·64	590	800	4	1·3
$Gd_2(SO_4)_3 . 8H_2O$	0·57	0·21	1380	1600		
$Cu(NH_4)_2(SO_4)_2 . 6H_2O$	9·1	4·5	200	85	6	4·5
$CuSO_4 . 5H_2O$	6·7	2·4	370	50	$\simeq 7$	25

Units of τ_S are 10^{-10} sec. H_l = internal field; H_p = perpendicular magnetic field to halve absorption. A_1 = exchange constant to account for difference of experimental and theoretical values of τ_S; A_2 = exchange constant calculated from specific heat data.

§ 5. EXPERIMENTAL METHODS

Paramagnetic relaxation phenomena in an alternating magnetic field can be described in terms of a complex susceptibility $\chi = \chi' - i\chi''$ which will depend on temperature, frequency of the alternating field, and on any steady field superposed on the alternating field. Experimental methods can conveniently be classified as methods for the measurement of χ' (dispersion measurements) and methods for χ'' (absorption measurements). Measurements of χ' can be made by observing the variation of the self-inductance or the mutual inductance of a coil system surrounding a specimen of paramagnetic salt. Measurements of χ'' can be made either calorimetrically, from the absorption of heat in the specimen, or electrically, from the losses in a coil system surrounding the specimen. The spin–lattice relaxation times of the order of 0·01 second occurring at liquid helium temperatures can be investigated by audio-frequency measurements. An A.C. balance method of measuring the mutual inductance and losses in a coil system round the specimen serves to determine χ' and χ'' simultaneously. Measurements at higher temperatures require greater sensitivity, since the susceptibility of the salt is smaller, and higher frequencies of measurement. Measurements of χ' at radio frequencies are made by determining the effect of the salt on the resonant frequency of a tuned circuit, using a heterodyne method to increase the sensitivity of measurement. A still more sensitive method of measurement is the determination of χ'' from the heating of the salt in a strong radio-frequency field.

Absorption Measurements at Radio Frequencies

This was the method used in the first experiments in which paramagnetic relaxation effects were detected (Gorter 1936 a, b). The technique used has since been improved in detail, but the principle of the method remains unaltered. If a specimen of paramagnetic salt placed in an alternating magnetic field of amplitude H_1 has a susceptibility with a component χ'' out of phase with the alternating field, the energy absorbed per gramme per second is

$$A = \tfrac{1}{2}\omega\chi'' H_1^2. \qquad \ldots\ldots(21)$$

Measurement of the heat generated in the salt then gives a measure of χ''.

Gorter in his first experiments used an oscillator working at frequencies of 10 to 27 Mc/s. giving fields at the specimen up to 8 oersteds. In later experiments the field strength and the frequency were raised. Dijkstra, Gorter and Van Paemel (1942) used a 10 kw. oscillator at frequencies up to 10 Mc/s. Volger, de Vrijer and Gorter (1947 a, b, c) used a 500 w. oscillator giving fields up to 3 oersteds at 78 Mc/s. In all the experiments the coil giving the alternating field surrounded a Dewar vessel of liquid nitrogen or liquid hydrogen. The specimen itself was contained in a thin-walled glass tube suspended in a glass vacuum enclosure by a long capillary tube connecting it to a sensitive manometer. The tube containing the specimen served both as a calorimeter and as the bulb of a constant volume gas thermometer by which the rise in temperature on applying an oscillatory field could be measured. It was of course essential to avoid the presence of any metallic parts in the calorimeter.

In an experiment, the specimen was first cooled to the temperature at which measurements were to be made, by pumping the liquid in the Dewar vessel. The space surrounding the calorimeter was then evacuated, and the temperature measured at fixed intervals of time to establish the rate of drift. The oscillatory

field was then switched on for a few minutes and the rise in temperature measured. The experiments could be made at different temperatures and frequencies, and with steady magnetic fields applied parallel or perpendicular to the oscillatory field.

In order to convert the measured rate of rise of gas thermometer pressure into values of χ'', the heat capacity of the specimen and calorimeter and the amplitude of the oscillatory field must be known. It is not easy to make absolute determinations of these quantities, and in spin–lattice relaxation measurements the usual practice has been to fit the results (expressed in arbitrary units) of measurements at different temperatures, frequencies and magnetic fields, to the Casimir–Du Pré formulae (equations (6), (7) and (8)) a correction being made for the absorption due to spin–spin relaxation. The process involves the adjustment of a rather large number of variables, but the results obtained for the relaxation times are strikingly consistent, while the specific heats determined in this way are in good agreement with the results of measurements by other methods.

Spin–spin relaxation times, on the other hand, can be determined only from absolute measurements of the absorption, using the equation

$$A_S = \tfrac{1}{2}\omega^2 \chi_S \tau_S H_1^2. \qquad \ldots\ldots(11)$$

In their measurements at 78 Mc/s. Volger, de Vrijer and Gorter calibrated the field strength of their oscillatory field by measuring the torque on a wire loop suspended at an angle to the field, on the principle of Fleming's dynamometer (Glazebrook 1922). To convert the gas thermometer measurements into heat input, the effect of known amounts of heat was found by induction heating, a copper tube being fixed round the calorimeter and heated by eddy currents at 50 c/s. An important correction to the R.F. heat input is the dielectric heating of the specimen and calorimeter, which becomes appreciable at high radio frequencies. One method of determining the correction is to measure the heating when the paramagnetic salt is replaced by a similar diamagnetic salt; another is to measure the absorption when a very large steady field is applied perpendicular to the oscillatory field. The spin–spin absorption is then greatly reduced, and the remaining heating can be ascribed to dielectric absorption.

Dispersion Measurements

This method of measurement was used in the first experiments on spin-lattice relaxation effects (Gorter and Brons 1937). A specimen of paramagnetic salt was inserted into a coil forming part of a radio-frequency tuned circuit, and the change of resonant frequency was measured by a heterodyne method. The change gives the increase in self-inductance of the coil, which is proportional to the in-phase component of susceptibility of the salt χ'. (The out-of-phase component χ'' changes the damping of the circuit.)

It is found that in the absence of any steady magnetic field the increase in the inductance, and hence the susceptibility of the salt, is independent of the frequency of the oscillatory field. By taking the susceptibility under these conditions as equal to the static susceptibility χ_T of the salt, which is known, these measurements serve to calibrate the apparatus, the results of other experiments being expressed as the ratio χ'/χ_T, that is, the ratio of the increase of inductance at a given steady field to the increase in the absence of a steady field, at the same frequency and temperature. With a steady magnetic field applied parallel to the oscillatory field, the susceptibility becomes a function of frequency, diminishing as the

frequency is raised, until at sufficiently high frequencies it reaches a new value independent of frequency—the adiabatic value χ_S. The variation has the form predicted by the Casimir–Du Pré theory, and the experiments can therefore be interpreted to give the spin–lattice relaxation time τ, and the ratio χ_S/χ_T from which the magnetic specific heat can be obtained.

Extensive measurements have been made at Groningen and at Amsterdam by the dispersion method. The most advanced technique was used by Broer (Broer and Schering 1943, Broer 1947, Broer and Kemperman 1947). In these experiments, made at frequencies from 100 kc/s. to 14 Mc/s., the measuring coil was suspended inside a screening can in a Dewar vessel containing liquid nitrogen. It formed part of the tuned circuit of an oscillator, whose frequency could be compared with a reference oscillator of fixed frequency by mixing the output of the two and detecting the audio difference frequency. A large water-cooled solenoid surrounding the Dewar vessel produced constant magnetic fields up to 3,000 oersteds parallel to the oscillatory field.

Since all the measurements are electrical, the dispersion method has the advantage over the absorption method in that a large number of specimens can be investigated without dismantling the apparatus, each specimen of paramagnetic salt, enclosed in a thin-walled glass tube, being lowered into the measuring coil as required. The chief difficulties arise from dielectric and damping effects. It is essential that the effect of inserting the salt should be purely magnetic, and not electric. To ensure that the salt should not affect the self-capacity of the coil, Broer and Schering (1943) screened the specimen from the coil by a mat of insulated copper wires. The remaining dielectric effect was estimated by inserting a diamagnetic specimen. To ensure that the damping of the circuit on introducing the specimen did not affect the tuning, they constructed an oscillator with an auxiliary circuit by which the tuned circuit could be damped electronically by a controllable amount, and kept the total damping always constant.

Besides the measurements made in Holland, those of Starr made at M.I.T. (Starr 1941) are of special interest, since he was able to work with steady fields up to 60,000 oersteds using the M.I.T. high power solenoid. His results for χ_S/χ_T are in agreement with the Dutch results, but there are some differences in the estimates of relaxation times.

Measurements at Liquid Helium Temperatures

The method used for audio-frequency measurements at liquid helium temperatures was devised by Casimir, Bijl and Du Pré (1941). A specimen of paramagnetic salt, cooled by liquid helium, is surrounded by a mutual inductance system, the primary of which carries the oscillatory magnetizing current, and by a solenoid giving a large steady field parallel to the oscillatory field. Changes in the susceptibility of the salt are measured by their effect on the mutual inductance system. The induced voltage in the secondary due to the component of magnetization of the specimen in phase with the oscillatory field is balanced by a decade mutual inductance having its primary and secondary in series with those of the coil surrounding the specimen. The difference ΔM between the value of the compensating inductance M and its zero value in the absence of the specimen gives χ' in arbitrary units. The secondary voltage due to the out-of-phase component of magnetization is balanced by a potentiometer system by which the voltage across a resistance in the primary circuit can be injected into the secondary

circuit. If the difference of this resistance from its zero value is ΔR at a frequency $\omega/2\pi$, $\Delta R/\omega$ gives χ'' in the same units as χ'. To convert the arbitrary units into absolute values, use is made of the fact that in the absence of a steady field, $\chi''=0$ and $\chi'=\chi_T$, the static susceptibility of the salt, which is known from other experiments. Since it is not usually possible to remove the salt from the apparatus during an experiment, the zero value of the mutual inductance M is found by measuring its value with the salt at different known temperatures, and plotting M against $1/T$. This gives a straight line graph, as the salts used closely follow Curie's law, and extrapolation to $1/T=0$ gives the zero value of M.

In the experiments of Benzie and Cooke (1950 a, b), whose apparatus is described here (Figure 2), the same method was followed. The specimen, in the form of small crystals, was enclosed in a perforated container of Perspex, the inside of which was spheroidal, to allow demagnetizing corrections to be applied

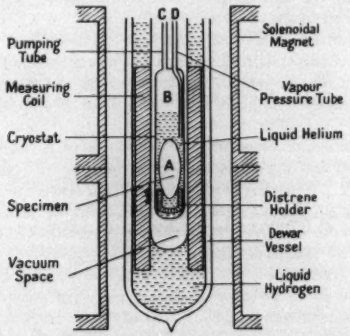


Figure 2. Apparatus for spin–lattice relaxation measurements at liquid helium temperatures (Benzie and Cooke).

to the calculation of the susceptibility of the salt. The container stood in a bath of liquid helium in a glass cryostat, which was suspended in a glass vacuum case surrounded by liquid hydrogen in a Dewar vessel. The temperature of the helium bath was measured by a helium vapour pressure thermometer, and could be lowered by evaporation to just below 1° K. The mutual inductance coil by which the susceptibility of the salt was measured was immersed in the liquid hydrogen, which formed an effective thermostat, and a solenoid giving steady fields up to 1,000 oersteds surrounded the Dewar vessel. One of the difficulties of the measurements was that eddy currents induced in the solenoid circuit by the alternating field would affect the mutual inductance of the measuring coil, so that changes in the eddy currents due, for instance, to changes in the resistance of the solenoid circuit would give rise to spurious changes in the apparent susceptibilities. This was avoided by using a mutual inductance system with a very small external field (Casimir, Bijl and Du Pré 1941). Balance of the measuring mutual inductance and its compensating circuit was indicated by a null detector, consisting of a tuned valve amplifier with an input sensitivity of 1 μv., feeding the Y-plates of a cathode-ray oscillograph. As the measurements were of course highly frequency sensitive, the frequency of the alternating field was standardized against a constant

frequency source connected to the X-plates. Each experiment consisted of a series of calibration measurements made in zero steady field with the salt at various known temperatures from 90° K to 1° K., followed by measurements at different temperatures and frequencies with a steady magnetic field applied parallel to the oscillatory field. The first series served to establish the zero value of the mutual inductance and the exact temperature dependence of the susceptibility, the second determined the relaxation effects. The frequencies and magnetic fields used varied of course from one salt to another, in order to cover a range sufficient for the determination of the spin–lattice relaxation time and the adiabatic susceptibility.

§ 6. SPECIFIC HEATS

All the experimental methods described above may be applied to the measurement of magnetic specific heats. Because of its magnetic properties, a paramagnetic salt possesses a specific heat additional to the specific heat of thermal vibration possessed by all solids. Interactions of the magnetic ions with one another and with their surroundings cause a separation of energy levels which in other substances are completely degenerate. For example, in chromium caesium alum, a salt whose energy levels are known from spectroscopic measurements (Bleaney 1949), the Cr^{+++} ion has $S=3/2$ giving 4 spin energy levels. These are separated by the Stark effect of the electric field of the crystalline lattice into two pairs, with an energy difference $\Delta E=k\theta$, where $\theta=0{\cdot}19°$. (Other interactions cause a further spreading of these levels, but for the moment we consider only the Stark effect.) At high temperatures ($T\gg\theta$) all the levels are equally occupied; at very low temperatures ($T\ll\theta$) only the lower pair is occupied. Corresponding to this change in the population of the energy levels there is an anomaly in the specific heat with a maximum in the neighbourhood of $T=\theta$, tailing off at high temperatures to a specific heat $C/R=\frac{1}{4}(\theta/T)^2$. The measured magnetic specific heat of chromium caesium alum is given by $CT^2/R=0{\cdot}011$, of which the Stark effect contributes 0·010.

Except at temperatures near the maximum of the anomaly, the magnetic specific heat is generally too small in comparison with the ordinary lattice specific heat to be determined by direct calorimetry. Relaxation methods, however, allow it to be determined indirectly, since they afford a method of separating it from the lattice specific heat. The adiabatic susceptibility χ_S which can be found by relaxation methods is not of course the susceptibility of the isolated salt, but the susceptibility when thermal contact between the magnetic ions and the lattice is negligible. Correspondingly the specific heats C_H and C_M entering in the equations

$$\chi_T/\chi_S=C_H/C_M \qquad \text{.....(3)}$$

and

$$C_H-C_M=T\left(\frac{\partial M}{\partial T}\right)_H^2\frac{1}{\chi_T} \qquad \text{.....(4)}$$

are the specific heats of the magnetic spin system. The isothermal susceptibility χ_T is usually known from standard susceptibility measurements (e.g. Volger, de Vrijer and Gorter 1947 a); measurement of the ratio χ_S/χ_T by experiments on spin–lattice relaxation then allows the specific heats to be calculated. From equations (3) and (4) we have

$$C_M=T\left(\frac{\partial M}{\partial T}\right)_H^2\frac{\chi_S}{\chi_T(\chi_T-\chi_S)}. \qquad \text{.....(22)}$$

The specific heat at constant magnetization C_M is generally of more interest than C_H, since for an ideal paramagnetic, i.e. one for which $M=f(H/T)$, C_M is independent of the magnetic field at a given temperature. The paramagnetic salts discussed here approximate to ideal paramagnetics, and at high temperatures can be taken to obey Curie's law : $M=CH/T$. In this case, equation (22) becomes

$$\frac{C_M T^2}{C} = \frac{H^2 \chi_S}{\chi_T - \chi_S}. \qquad \ldots\ldots(23)$$

At liquid helium temperatures, two factors cause departures from Curie's law. Interactions of the magnetic ions may cause the salt to obey a Curie–Weiss law rather than the simple Curie law, and magnetic saturation may be appreciable in the fields used for spin–lattice relaxation measurements. The specific heat must then be calculated from the thermodynamic formula (22), instead of the simpler equation (23). The form $C_M T^2/C$ is a convenient method of expressing the results of specific heat measurements, since it is closely related to measured quantities and is independent of temperature at temperatures high compared with that at which the specific heat passes through its maximum. Gorter (1947) gives specific heat results in this form, writing $C_M T^2 = b$ and giving values of b/C. His tables include data from various doctorate theses, which are not generally available. Since this publication, relaxation measurements of specific heats have been published by Broer (1947) on some chromium salts and iron ammonium alum, by Broer and Kemperman (1947) on some copper and silver salts, by Kramers, Bijl and Gorter (1950) on iron ammonium alum and potassium chrome alum, and by Benzie and Cooke (1949, 1950 b) on salts of copper, iron, manganese, chromium, nickel and gadolinium.

Interpretation of Magnetic Specific Heats

The chief interest of the magnetic specific heat measurements lies in the information they give on the energy levels of the magnetic ions, from which, for example, their magnetic and thermal properties at the very low temperatures attainable by the magnetic method of cooling can be calculated. Until recently it was supposed that the only important factors determining the energy levels were the Stark effect of the electric field of the crystalline lattice, and the magnetic dipole interaction of the spins. The magnetic dipole interaction was shown by Van Vleck (1937) to contribute to the specific heats an amount C_{mag} given by

$$\frac{C_{mag} T^2}{C} = \tfrac{1}{2} H_i^2. \qquad \ldots\ldots(24)$$

This can therefore be calculated if the crystal structure and susceptibility of the salt are known, and by attributing the remainder to the Stark effect the Stark separation of the energy levels could be calculated (Hebb and Purcell 1937, Cooke 1949). Recently, however, it has been discovered that two factors may contribute appreciably to the specific heats. Gorter (1948) pointed out that in ions having both a spin moment and a nuclear magnetic moment, the coupling of the two may produce an appreciable splitting of the spin energy levels. Further, exchange interaction between the spins may be appreciable (Opechowski 1937, 1948, Gorter and Van Vleck 1947, Pryce 1948). The Stark splitting of the energy levels cannot therefore be obtained so readily. It is, however, possible to analyse the specific heats more exactly by measurements on diluted paramagnetic salts—that is, mixed crystals of a paramagnetic and a similar diamagnetic salt. The

magnetic dipole and exchange interactions are, of course, diminished by dilution, and if the distribution of paramagnetic and diamagnetic ions in the mixed crystal is random, their contribution to b/C may be expected to be proportional to the concentration of paramagnetic salt. Extrapolation of the results to zero concentration then allows the specific heat to be analysed into a part dependent on the concentration (magnetic and exchange interactions) and a part independent of concentration (Stark splitting and nuclear hyperfine splitting of the energy levels). The exchange contribution to the specific heat can be found by subtraction of the magnetic specific heat, which can be calculated, from the concentration-dependent part. The Stark and hyperfine splittings cannot be resolved by specific heat measurements of the kind discussed here.

This technique has been used (Benzie and Cooke 1949) in experiments on cupric double sulphates diluted with the corresponding zinc salts. In these salts the Cu^{++} ion is effectively in a state with $S = 1/2$, and so has only a two-fold degeneracy. A well-known theorem of Kramers (1930) states that this is not resolved by an electric field. The cupric salts therefore present a simple case in which the whole of the specific heat independent of concentration is due to the nuclear splitting. The results of the measurements are shown in Table 2.

Table 2. Analysis of Specific Heats of Copper Tutton Salts

Salt	$\frac{C_M T^2}{R}$	$\frac{C_{nuc} T^2}{R}$	$\frac{C_{mag} T^2}{R}$	$\frac{C_{ex} T^2}{R}$
Copper ammonium sulphate	8·7	1·1	1·3	6·3
Copper potassium sulphate	6·0	1·1	1·3	3·6
Copper rubidium sulphate	3·3	1·1	1·3	0·9
Copper caesium sulphate	2·6	1·1	1·2	0·3

Units of CT^2/R are $(\text{degrees})^2 \times 10^{-4}$.

It is seen that whereas the nuclear specific heat and the (calculated) magnetic dipole interaction specific heats are practically the same for all the salts, the exchange interaction contributes widely different amounts, and in two of the salts is by far the largest effect. Similar results, not yet published, have shown that exchange interaction is appreciable in cobalt ammonium sulphate and manganous ammonium sulphate. It seems certain therefore that earlier estimates of the magnetic energy level separations in paramagnetic salts, based on measurements of the specific heats of the pure salts, are not dependable.

REFERENCES

Bagguley, D. M. S., Bleaney, B., Griffiths, J. H. E., Penrose, R. P., and Plumpton, B.I., 1948, *Proc. Phys. Soc.*, **61**, 542.

Benzie, R. J., and Cooke, A. H., 1949, *Nature, Lond.*, **164**, 837; 1950 a, *Proc. Phys. Soc.* A, **63**, 201; 1950 b, *Ibid.*, **63**, 213.

Bijl, D., 1941, *Physica*, **8**, 497; 1950, *Proc. Phys. Soc.* A, **63**, 405.

Bleaney, B., 1949, *Phys. Rev.*, **75**, 1962; 1950, *Proc. Phys. Soc.* A, **63**, 407.

Broer, L. J. F., 1943, *Physica*, **10**, 801; 1945, *Thesis*, Amsterdam; 1947, *Physica*, **13**, 353.

Broer, L. J. F., and Kemperman, J., 1947, *Physica*, **13**, 465.

Broer, L. J. F., and Schering, D. C., 1943, *Physica*, **10**, 631.

Casimir, H. B. G., Bijl, D., and Du Pré, F. K., 1941, *Physica*, **8**, 449.

Casimir, H. B. G., and Du Pré, F. K., 1938, *Physica*, **4**, 579.

Cooke, A. H., 1949, *Proc. Phys. Soc.* A, **62**, 269.

Debye, P., 1938, *Phys. Z.*, **39**, 616.

DIJKSTRA, L. J., GORTER, C. J., and VAN PAEMEL, O., 1942, *Physica*, **9**, 673.
FIERZ, M., 1938, *Physica*, **5**, 433.
GLAZEBROOK, R., 1922, *A Dictionary of Applied Physics* (London : Macmillan & Co.).
GORTER, C. J., 1936 a, *Physica*, **3**, 503; 1936 b, *Ibid.*, **3**, 1006; 1947, *Paramagnetic Resonance* (Amsterdam : Elsevier Publishing Co.); 1948, *Physica*, **14**, 504.
GORTER, C. J., and BRONS, F., 1937, *Physica*, **4**, 579.
GORTER, C. J., and KRONIG, R. DE L., 1936, *Physica*, **3**, 1009.
GORTER, C. J., TEUNISSEN, P., and DIJKSTRA, L. J., 1938, *Physica*, **5**, 1013.
GORTER, C. J., and VAN VLECK, J. H., 1947, *Phys. Rev.*, **72**, 1128.
DE HAAS, W. J., and DU PRÉ, F. K., 1938, *Physica*, **5**, 969.
HEBB, M. H., and PURCELL, E. M., 1937, *J. Chem. Phys.*, **5**, 338.
HEITLER, W., and TELLER, E., 1936, *Proc. Roy. Soc.* A, **155**, 629.
KRAMERS, H. A., 1930, *Proc. K. Ned. Akad. Wet.*, **33**, 959.
KRAMERS, H. C., BIJL, D., and GORTER, C. J., 1950, *Physica*, **16**, 65.
KRONIG, R. DE L., 1939, *Physica*, **6**, 33.
OPECHOWSKI, W., 1937, *Physica*, **4**, 181; 1948, *Ibid.*, **14**, 237.
PRYCE, M. H. L., 1948, *Nature, Lond.*, **162**, 538.
STARR, C., 1941, *Phys. Rev.*, **60**, 241.
TEMPERLEY, H. N. V., 1939, *Proc. Camb. Phil. Soc.*, **35**, 256; 1947, *Ibid.*, **43**, 118.
TURNER, E. H., SACHS, A. M., and PURCELL, E. M., 1949, *Phys. Rev.*, **76**, 465.
VAN VLECK, J. H., 1937, *J. Chem. Phys.*, **5**, 320; 1940, *Phys. Rev.*, **57**, 426; 1948, *Ibid.*, **74**, 1168.
VOLGER, J., 1946, *Thesis*, Leiden.
VOLGER, J., DE VRIJER, F. W., and GORTER, C. J., 1947 a, *Physica*, **13**, 621; 1947 b, *Ibid.*, **13**, 635; 1947 c, *Ibid.*, **13**, 653.
DE VRIJER, F. W., and GORTER, C. J., 1948, *Physica*, **14**, 617.
WALLER, I., 1932, *Z. Phys.*, **79**, 370.
WRIGHT, A., 1949, *Phys. Rev.*, **76**, 1826.

CYCLIC ACCELERATORS

By J. H. FREMLIN and J. S. GOODEN
Birmingham University

CONTENTS

PAGE

§ 1. Introduction ... 295
§ 2. Fixed-frequency cyclotron ... 296
(i) General principles ... 296
(ii) Detailed consideration of the different regions of acceleration ... 298
(*a*) Ion source and initial conditions ... 298
(*b*) Conditions required during the main part of the acceleration if the maximum energy is to be achieved ... 302
(*c*) Conditions required for extraction of the beam ... 307
(iii) The magnet ... 309
(iv) The radio-frequency system ... 310
(*a*) The generator ... 310
(*b*) Construction of dees and resonant lines ... 312
(v) Limits of the fixed-frequency cyclotron ... 315
§ 3. The frequency-modulated cyclotron, or synchro-cyclotron ... 317
(i) Effect of frequency variation ... 317
(ii) Early experimental work ... 319
(iii) The present position of the synchro-cyclotron ... 321
§ 4. Betatrons ... 324
(i) Introduction ... 324
(ii) Principle of operation ... 325
(iii) Physical construction ... 327
(iv) Operation ... 330
(v) Future possibilities ... 331
§ 5. Electron synchrotron ... 331
(i) General principles ... 331
(ii) Physical construction ... 334
(iii) Operation ... 336
(iv) Future possibilities ... 336
§ 6. The proton synchrotron ... 337
(i) General principles ... 337
(ii) Physical construction ... 340
(iii) Performance and future prospects ... 345
§ 7. Conclusion ... 346
Acknowledgments ... 347
References ... 348

§ 1. INTRODUCTION

In the period following the end of the war immense sums of money have been invested in particle accelerators. The historical development of our present interest in beams of fast particles has been well described more than once and hardly needs any repetition here. The functions for which high-energy machines have been designed may be divided into two main groups: those for direct investigations of particulate properties, and those designed as 'factories' for the production of such consumer goods as radio-isotopes for medical research or hard x-rays for the inspection of heavy metal structures. The separation in principle seems worth keeping in mind in spite of the fact that many existing and

prospective machines have been intentionally designed for use in both fields. The distinction matters partly because a dual-purpose design should not be undertaken without careful consideration, as the necessary compromises may have a serious effect on the peak performance possible in one or both fields. Still more it matters because while a machine in the first group is, however large, a piece of physics apparatus, one in the second group is essentially an engineering unit. The fact that the latter units are still largely built and run by physicists in physics departments is due merely to the inevitable time-lag in the production of the appropriate kinds of electrical engineers. A discussion of fixed-frequency cyclotrons, for example, will presumably become in time as inappropriate for publication in these *Reports* as a discussion of the design of high-power alternators would be at present.

While physicists still retain some influence on accelerator-design, it seems to be worth while to describe and compare the principles and capabilities of the many kinds of accelerator now under development. Excellent papers have recently been published on electrostatic accelerators by Van de Graaff *et al.* (1947), and on linear accelerators by Fry and Walkinshaw (1949), and the design of high-voltage rectifier units has already gone far along the path towards adoption by the engineers.

We shall limit ourselves therefore to the more important devices requiring the use of magnetic fields to control the particle orbits. After separate consideration of each of the main types of accelerator we shall in a final section discuss briefly the range of operation for which each machine is economically best suited.

§ 2. FIXED-FREQUENCY CYCLOTRON

The cyclotron is the oldest of the particle accelerators with which we shall be concerned, the first one having been successfully constructed twenty years ago by Professor E. O. Lawrence and Dr. M. S. Livingston, who thus laid the foundations of one of the major fields of modern physics (Lawrence and Livingston 1931). Since this early beginning, detailed descriptions of individual machines have been published in many places. It is considered, therefore, that the best approach will be to explain the principles of operation as they are now understood, with particular reference to those effects which are neglected in the elementary theory but which nevertheless are responsible for much of the detailed design and nearly all of the limitations of modern fixed-frequency cyclotrons. Typical recent designs of different parts of the equipment will be described but no attempt will be made to give any complete survey of the large number of different constructions now in practical use.

(i) *General Principles*

A particle of mass M and charge e moving in a magnetic field H with velocity v will follow a circular orbit of radius r given by

$$r = \frac{Mv}{He}. \qquad \ldots\ldots(1)$$

The periodic time will be given by

$$\tau = \frac{2\pi r}{v} = \frac{2\pi M}{He}$$

and the angular frequency of rotation in the orbit, ω_p, will be

$$\omega_p = \frac{He}{M}. \qquad \ldots\ldots(2)$$

The frequency is then independent of the particle velocity at speeds sufficiently low for M to be nearly constant. To show the order of magnitude of energy obtainable with this restriction, a 1% increase of mass represents about 10 MeV. for a proton and 20 MeV. for a deuteron, but only 5 kev. for an electron. The fact that the rotation frequency is independent of velocity represents the basic requirement of the cyclotron. From the figures just given, this kind of accelerator is therefore suitable only for heavy particles. The principle, briefly, is as follows.

In a uniform and constant magnetic field an alternating electric field of constant angular frequency ω_E is used to accelerate particles, the position of which in their orbit bears a constant relation to the phase of the electric field. The essential components of a cyclotron are shown in Figure 1. Two hollow dee-shaped

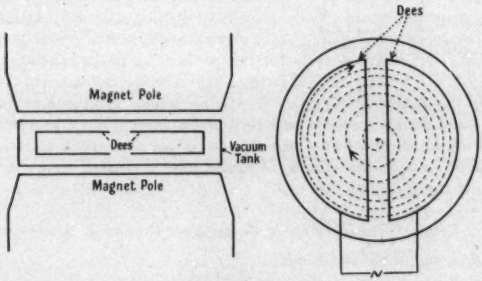


Figure 1.

electrodes, connected to a high-frequency voltage supply, are mounted in a vacuum box in a strong magnetic field perpendicular to the planes of symmetry of the dees. Then a charged particle initially at rest at the centre of symmetry of the system will follow a path of the form shown by the dotted line if $\omega_E = \omega_p$ (see equation (2)).

If the extension of the electric field between the dee-edges is small, the particle will pick up an energy $eV\cos\phi$ at each transit of the gap, and $2eV\cos\phi$ in each complete revolution, where $V\cos\omega_E t$ is the instantaneous voltage difference between the dees and $\omega_E t = \phi$ gives the phase of the electric field at which the particle starts.

The frequency relation given in equation (2) shows that the electric field will be exactly reversed while the particle travels between one gap-transit and the next, so that it is similarly accelerated on each occasion. Each dee acts in effect as a Faraday cage with no internal high-frequency field, so that except when crossing a gap the particle is acted on only by the magnetic force and pursues a semicircular path. After N complete revolutions the particle will have total energy $eV\cos\phi(2N-k)$, where k represents the fact that the full gap-energy may

not be obtained in the first revolution or so, when the orbit is comparable in size to the dee gap. Then, since we are at this stage limiting ourselves to the non-relativistic region,

$$\tfrac{1}{2}Mv^2 = eV(2N-k)\cos\phi,$$

so that

$$v = \left(\frac{2e}{M}\right)^{1/2}\{(2N-k)V\cos\phi\}^{1/2}.$$

From (1)

$$r_N = \frac{1}{H}\left(\frac{2M}{e}\right)^{1/2}\{(2N-k)V\cos\phi\}^{1/2}. \qquad \ldots\ldots(3)$$

Hence

$$\tfrac{1}{2}Mv^2 = \frac{e^2H^2r_N^2}{2M} \qquad \ldots\ldots(4)$$

and

$$N_{\max} \simeq N_{\max} - \tfrac{1}{2}k = \frac{E_{\max}}{2V\cos\phi},$$

where $E_{\max}$ electron volts is the maximum energy obtained.

Equations (3) and (4) show first that as the energy and radius increase the orbits are more and more closely crowded together, and second that the final energy depends only on H and $r_{\max}$ for a given particle, as is of course clear from (1).

The first requirement of cyclotron design then, for the production of particles of any desired energy, is a magnet giving a field of sufficient extent and magnitude to provide the required $(Hr)_{\max}$. A particle whose orbit reaches the design radius $r_{\max}$ will then necessarily have the correct energy, but a number of conditions must be satisfied before a particle can in fact reach such a radius. These conditions lie outside the simple idealized case discussed above and will be dealt with in the following sections.

(ii) *Detailed Consideration of the different Regions of Acceleration*

(*a*) *Ion source and initial conditions.*

In the earliest machines a column of ions was formed along the axis of symmetry of the system by electrons from a hot filament held near the floor of the vacuum tank and maintained at 200 volts or so negative with respect to the tank. The electron beam followed closely the lines of magnetic field until it was collected on an anticathode of graphite or other heat-resistant material or directly on the tank cover itself. This method is simple but requires the ionization to take place at the prevailing tank pressure, which is thus subject to conflicting requirements; if it be high there is a risk of spark-over from dee to tank, but if low there is a poor supply of ions.

In consequence it is usual nowadays to use one form or another of gas-arc source, in which gas is fed to an enclosed space containing the heated filament, and ions are allowed to diffuse out through a hole with sufficiently high impedance to gas flow to maintain a useful pressure difference between the filament region and the main tank (Livingston 1946). In Figure 2 are shown various recent forms of arc source (McMillan and Salisbury 1939, Atterling 1948). Important features are that the arc body must be extremely well cooled as it suffers serious back-bombardment from ions which are initially caught, but not held, by the high-frequency field, and that the ions are permitted to emerge only over the region from which they can be successfully accelerated to the target or exit-port. A convenient limitation is obtained by the use of a 'hooded' arc (originated in the

U.S.A. during the war: Cowie and Ksanda 1945) as in Figure 2 (*b*). With this a high concentration of ions can be obtained in a closely defined position. To get the best results the source must be mechanically adjustable, preferably during operation, as it is not practicable to calculate completely the full paths of an adequately representative set of the accelerated particles.

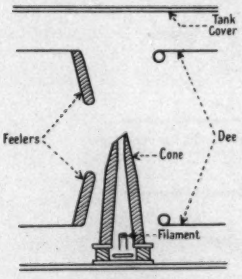


Figure 2 (*a*). University of California ion source, 1939.

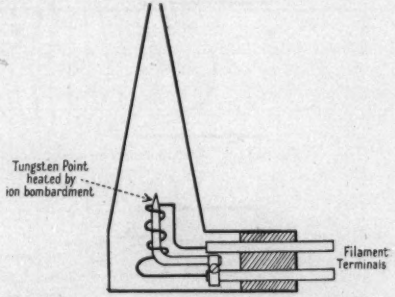


Figure 2 (*c*). Proposed ion source for Nobel Institute for Physics, Stockholm.

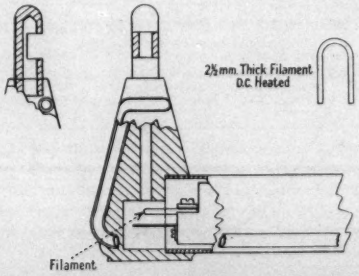


Figure 2 (*b*). University of Birmingham ion source, 1949.

Although complete calculation would be unprofitable, it is desirable to understand something of the behaviour of the ions near the source. Conditions differ in several important particulars from those of the idealized case. The accelerating electric field extends over a considerable distance around the source in every direction and is continuously effective—not merely at the instant when the particles are crossing the centre of the gap. The particles start, with finite initial velocities, from a region of finite size containing space charge. Finally the magnetic field is not necessarily exactly $\omega_E M/e$.

The effects of real conditions will be considered to a first order only, i.e. it will be assumed that they can be treated independently and with appropriately limited accuracy.

In Figure 3 are shown the instantaneous distributions of potential and of electric field between the dee edges of a typical cyclotron in the absence of the ion source. The distribution is little affected by the presence of the source, and it will be seen that the initially emergent ions moving close to the source are in a spatially nearly uniform field, $(V/l)\cos\omega t$ say, as indicated in Figure 4, where l is a

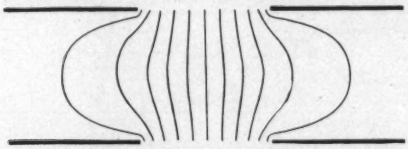


Figure 3 (*a*). Distribution of equipotential lines between the dee edges.

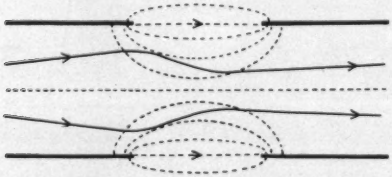


Figure 3 (*b*). Lines of electric force (dotted) and the (exaggerated) forms of path of particles between the dee edges.

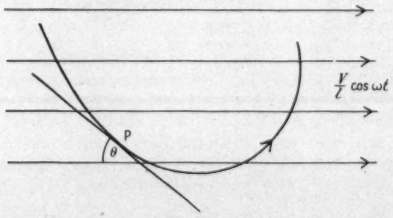


Figure 4.

distance somewhat greater than d, the distance between the dee edges. We shall deal first with particles moving in the median plane of the system. For these, when the instantaneous position of the particle is P (Figure 4), moving at an angle θ to the lines of force, the forces acting on the particles are $(eV/l)\cos\omega_R t\cos\theta$ along its path and $HeV+(eV/l)\cos\omega_R t\sin\theta$ perpendicular to its path. Assuming the electric field to be thus uniform, Mr. H. I. S. Allwood has calculated the initial paths of particles in the median plane of the Birmingham 60-inch cyclotron for a number of phases of entry into the field under representative working conditions (Allwood 1948). These paths are shown in Figure 5.

The most important point demonstrated by these is the tendency of particles starting with different phases to 'bunch' together with the particles of optimum

phase. The existence of this bunching had also been pointed out by Bohm and Foldy (1947). This leads to a reduced number of revolutions for particles entering the field at other phases, which, as will be seen below, affects in a number of different ways their chances of gaining the full energy.

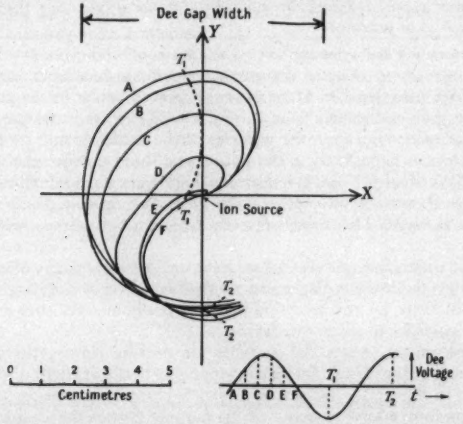


Figure 5. Initial ion paths in the cyclotron calculated by Mr. H. I. S. Allwood for deuterons in the Birmingham cyclotron assuming a peak dee-to-dee voltage of 200 kv. The lower figure shows the times of emission of the ions whose paths are shown. The dotted lines in the upper figure show the loci of ions at the instants T_1 and T_2.

The incomplete uniformity of field distribution and slight differences of magnetic field from the value $\omega_R M/e$ are very unlikely to affect this conclusion qualitatively. Of less importance than the bunching, but of some practical value, is the fact that from tracks of the type shown in Figure 5 we can make an estimate of the permissible dimensions of the ion-source body for any value of dee voltage.

Now let us consider the behaviour of particles whose paths are not limited to the median plane, owing either to an initial velocity v_z perpendicular to this plane or to an initial displacement z from it. So far we have considered no forces normal to the plane. If there were none, the initial displacement could be equal to the half-aperture of the dees, but the initial velocity v_z would have to be so small that a particle starting near the bottom of the dees, say, did not rise as much as the dee aperture in the whole acceleration time. This means that v_z must be below about 2×10^6 cm/sec., corresponding to a kinetic energy of 3–4 ev. for a deuteron. Since nearly all the ions emerging from the arc will have larger random energies than this, some restoring force acting towards the median plane will be most desirable. As can be seen from Figure 3 (*b*), the R.F. field will produce considerable forces normal to the median plane near the dee edges, and if the main magnetic field is not uniform this also will introduce such forces. These points will be

considered further in the discussion of the conditions controlling ion stability during acceleration.

A full discussion of the practical performance of the various types of ion source has been given by Livingston (1946).

(*b*) *Conditions required during the main part of the acceleration if the maximum energy is to be achieved.*

The elementary theory as applied to the main acceleration needs to be extended for two main reasons. Firstly, if more than an infinitesimal beam is required, we must consider particles with finite velocities perpendicular to the median plane and moving at finite distances from it; and secondly, we must take account of the fact that the motions of even the particles in the median plane are complicated by such factors as imperfectly uniform magnetic fields and the relativity-increase of mass. (We shall for convenience refer below to 'horizontal' and 'vertical' directions as measuring velocities parallel to and at right angles to the median plane respectively.) The theory of ion focusing has been given by R. R. Wilson (1940).

It will be shown that as the result of restoring forces the particles will normally oscillate about the median plane, and we shall begin by considering the stability of such oscillations, i.e. the condition that they shall not reach such an amplitude as to allow particles to strike the dees.

Near the source, where the particles are moving slowly, the main forces concerned will be the electric forces corresponding to the gradient of the potential lines shown in Figure 3. In fact, as a particle passes from one dee to the other, the two gaps form a kind of lens. If the particle crosses the centre of the field at the instant of maximum voltage, this lens is converging for particles moving parallel to the lens axis. This is because the particles are moving more slowly in the earlier, converging, part of the lens than in the later, diverging, part (see Figure 3(*b*)). If the field is decreasing as the particle crosses, the lens is even more strongly converging, as the diverging field beyond the centre line will be relatively weaker by the time the particle reaches it. On the other hand, if the field is *increasing* as the particle passes, the lens may become diverging in its net effect. In a practical system these time-dependent focusing effects are much stronger than the focusing effects on a particle passing through a static field.

Then if ϕ is the phase at which the particles cross the centre of a gap, measured from the instant when the gap-field is a maximum, the electric forces will be focusing if ϕ is positive, i.e. if the particle crosses the field *after* the field maximum. They will still be focusing if ϕ is very slightly negative, but will be strongly defocusing if ϕ is much negative, i.e. if the particle crosses the gap significantly before the field maximum. Unfortunately, as shown in Figure 5, all particles do in fact cross at or before the field maximum.

The magnitude of the forces is approximately proportional to the distance from the median plane, and hence, in the defocusing condition, if the particle phase remains constant, the distance of a given particle from the median plane will increase exponentially, or, with an initial supply of particles at all distances, the surviving current will decrease exponentially.

In the focusing condition the particles will oscillate about the median plane; and at constant phase, as the restoring force becomes weaker at larger particle

energies, the period and amplitude both increase as we go from the centre of the cyclotron outwards. The period of oscillation is much longer than the period of revolution of the particles, and the increase of the amplitude will be roughly as $\sqrt{r}$. Hence, if particles were available with positive phase of transit, the situation would be much better than when the gap forces are defocusing, but there would still be a serious loss of particles to the dees at large radii unless some further focusing were available.

Fortunately this can be done with the help of a controlled non-uniformity of magnetic field. In a field that decreases towards the outside the lines of force will be curved, as in Figure 11, so that away from the median plane there will be a radial component of field H_r. If the change of vertical component H_z with radius is $\partial H_z/\partial r$, then $\partial H_r/\partial z = -\partial H_z/\partial r$ since the tangential component $H_\theta = 0$ for a radially symmetrical field; and, as $H_r = 0$ at $z = 0$ (on the median plane), we can say to sufficient accuracy $H_r = -z\,\partial H_z/\partial r$. From Figure 11 it is clear that the resulting vertical force $-H_r ev$ on the particle will be towards the median plane for a field decreasing outwards, and away from the median plane for one increasing outwards.

It is useful to get an idea of the order of magnitude of the effects to be expected in either case. We have for the equation of vertical motion of a particle during its path

$$\frac{\partial^2 z}{\partial t^2} = -H_r ev = \frac{\partial H_z}{\partial r} zev,$$

v being the horizontal velocity. $\partial H_z/\partial r$ is positive for a defocusing field, negative for a focusing one. All the quantities except e are functions of r and hence of time, and a solution is not readily obtainable except in special cases, such for example as cases when $v\,\partial H_z/\partial r$ changes very slowly with time. The condition for simple harmonic motion in the oscillating case is that during the relevant part of the acceleration $\partial H_z/\partial r$ is falling off as $1/r$, and in this case a particular solution can easily be found for the correct total fall of H_z over any desired radial range, the period of oscillation then being

$$2\pi\left(\frac{\partial H_z}{\partial r} ev\right)^{-1/2} \quad \text{or} \quad \left(r\frac{\partial H_z}{\partial r}\frac{e}{2\pi\tau}\right)^{-1/2} = \tau_{\text{osc}} \text{ (say).} \quad \ldots\ldots(5)$$

If, for example, in the Birmingham 60-inch cyclotron, we consider a fall of 1% in H_z from a radius of 13 cm. (equal to the dee aperture) to 63 cm., we find that under working conditions we shall have an oscillation of period about fourteen times the period of revolution of the particles (or a frequency of about 8×10^5 c/s. against $11{\cdot}5 \times 10^6$ c/s.). The vertical amplitude is, according to the assumption, constant. In the real case $\partial H_z/\partial r$ does not fall off at first as r increases as assumed, and actually rises later. If we had a 1% *rise* in H_z, on similar assumptions, the displacement from the median plane would increase exponentially by a factor of at least 10^8 even for the most efficiently accelerated particles, and no reasonable change in assumed variation of $\partial H_z/\partial r$ would bring this down to practical limits, although in most cases it is probably less than 10^8. In Figure 6 (*a*) is shown the variation of the magnetic field H_z with radius in the Birmingham cyclotron, and in Figure 6 (*b*) the resulting value of H_r at 1 cm. from the median plane. In Figure 6 (*c*) is shown the corresponding restoring force measured in volts/cm. together with the total restoring force due to the combined electric and magnetic

focusing, acting on ions accelerated to full energy by the minimum possible dee-voltage, as calculated by Mr. W. I. B. Smith, to whom the authors are indebted for permission to publish these curves.

When $r\partial H_z/\partial r$ is not constant, we can still use the expression (5) above for the period of oscillation so long as it varies only slowly. It can be seen from Figure 6 (*c*) that the frequency of oscillation will be rising towards the outside and the amplitude will be falling. This makes permissible a reduction of dee aperture, with a resulting reduction in capacitance to earth. In Figure 7 is shown a beam

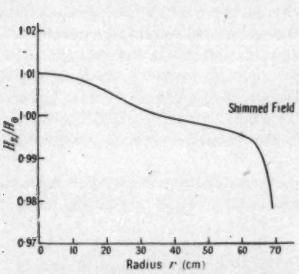


Figure 6 (*a*). Measured variation of magnetic field with radius—the Birmingham University cyclotron.

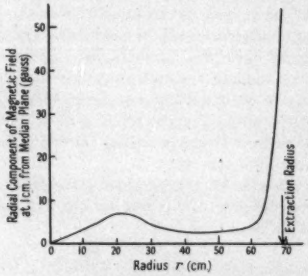


Figure 6 (*b*). Radial component of field in Birmingham University cyclotron.

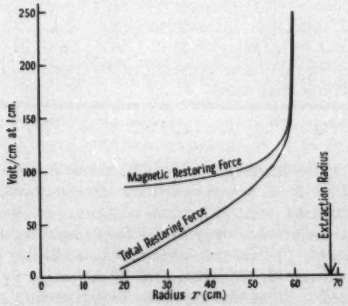


Figure 6 (*c*). Restoring forces in Birmingham University cyclotron.

profile experimentally found by Wilson (1938), together with a theoretical curve showing the oscillations about the median plane of a representative particle starting close to the dee surface in the region of the ion source.

We have seen that owing to the electric defocusing some positive magnetic focusing is essential, and hence the field *must* be made to decrease outwards, even if only slightly. This is unfortunate, as it would be desirable to *increase* the field at large radii to compensate for the effect of the relativity-increase of mass on the

period of revolution (see equation (2)). The best that can be done is to decrease the field as little as possible, confining the decrease to the radial region where focusing is most required. The vertical oscillations set up will then vary in frequency and amplitude under the combined effects of electric and magnetic focusing. The total vertical restoring force must be positive near the source to enable ions with some initial vertical velocity to be captured, and must be then kept of sufficient value to prevent the amplitude of oscillation from reaching the dee aperture. It might be thought that once the ions reached a good velocity no further focusing would be required, but it must be remembered that the vertical oscillation, though it may decrease in amplitude, is never damped out but actually has its *energy* increased, and that therefore a removal of the restoring forces would let the amplitude increase without limit. The phase of the vertical oscillation of different particles is of course quite random with respect to either radius or azimuth of the particles.

Before leaving this question it may be mentioned that a method of focusing has been proposed by Thomas by means of an azimuthally varying magnetic field which allows a rise in the *mean* value of the magnetic field (Thomas 1938). In a practical magnet, however, it is the *peak* magnetic field which is limited by the saturation properties of steel, rather than the mean value. Hence the very large

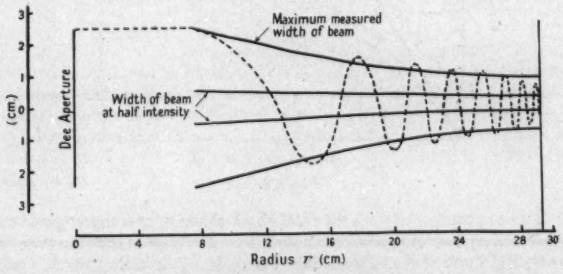


Figure 7. Calculation (dotted line) and measurement (full lines) on 37-inch Berkeley cyclotron by R. R. Wilson to show ion focusing.

azimuthal variation required to give a useful effect depresses the mean value possible over the whole dee system so much that the maximum energy obtainable is actually less instead of more for a given R.F. power and magnet size.

We can now return to consideration of the motions of the particles parallel to the median plane. We are faced with the fact that in order to get any useful proportion of current accelerated we must have a magnetic field falling towards the outside. The fall of magnetic field will then combine with the relativistic increase in mass to give an increase of period of revolution of the ions, while the accelerating R.F. voltage is of fixed frequency. For a given R.F. voltage the limiting energy possible will then be gained if the magnetic field is too high at the centre and too low outside. As was shown in §2(ii), in whatever phase a particle starts it will, after less than one cycle, be a little ahead in phase of the R.F. field. It will continue revolving too fast and will go further ahead in phase till it is nearly 90° ahead of the field. If the magnetic field is falling at the correct rate it will now be revolving at just the same rate as the alternations of R.F. field.

Further acceleration would cause it to revolve more slowly, and it would then lose in phase until it was 90° behind the field again. In practice the full range of phase is only possible if it is not desired to extract the beam. If the beam is to be extracted, the successive orbits must be reasonably well spaced out, which requires the phase to be a good deal less than 90° behind the field when the beam reaches the deflector plate. It should be noted that in most of the earlier period the particle crosses the gap ahead of the field maximum. Hence the electric forces are strongly defocusing and the need for good magnetic focusing is great.

Finally, to get a reasonable final current, particles over a fair phase-range must be accelerated. In Figure 8 is shown the variation with time of radius, and energy, for particles of various phases for a particular set of conditions.

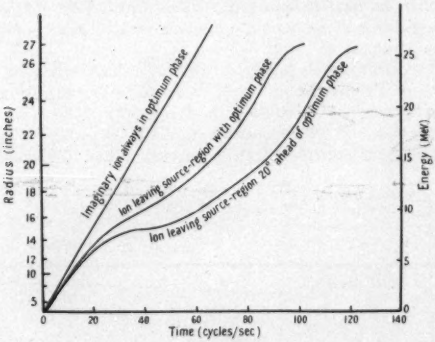

Radius (inches)
Energy (MeV)
Time (cycles/sec)
Imaginary ion always in optimum phase
Ion leaving source-region with optimum phase
Ion leaving source-region 20° ahead of optimum phase

Figure 8.

Since we are concerned with the *total* phase changes, and at any given radius there is a definite phase change *per cycle* given by the difference between the angular frequency of the R.F. field and the local value of He/M, it is clear that we gain by using high dee-voltages and hence a small number of revolutions. There will for each cyclotron be a critical voltage required for any given energy. For a useful current this critical voltage must be exceeded by perhaps 20–30%. The critical voltage increases as the square of the final energy E required. For a given design, in cyclotrons of varying size the overall power input will vary as $E^{7/2}$ (Livingston 1944), which shows very clearly that power requirements will form a very definite limitation to the possible range of use of fixed-frequency machines. This point will come up again in discussing the R.F. system of cyclotrons.

One further discrepancy between the motions as given by the simple theory and the real motions must be mentioned. If the R.F. voltage along the dees is not uniform, or if the magnetic field has azimuthal variation, the centre about which the particles move may itself move along a line or curve. Also, as a result of the inductance of the dee itself, an end-fed dee will have a higher voltage at the open end. This will lead to a migration of the centre of rotation towards the feeding end. In the Birmingham cyclotron this drift is about 1 cm., and in smaller machines it will usually be negligible.

The effect of non-uniformity of magnetic field may be seen if we suppose one half of the field to be stronger than the other, as in Figure 9 (*a*). The effective centre of rotation will then move along the dividing line between the strong and weak field in the direction in which the particles are moving in the weaker field, as the semicircles in this are the larger.

If, as is always the case, there is a radial change of field as well as either of these changes, the centre of rotation will precess about the centre of symmetry of the

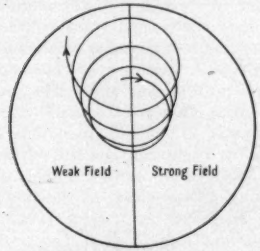


Figure 9 (*a*).

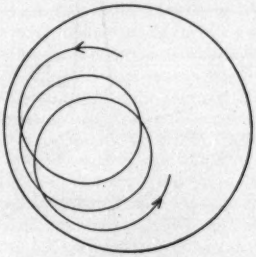


Figure 9 (*b*).

magnetic field, as shown in Figure 9 (*b*). The rate of precession can be controlled by shimming, and, as is shown in the next section, may be useful in decreasing the loss of particles in the extraction system. These effects are, however, most unhelpful in giving rise to a very large spread in energy and direction of the particles at a given point; the mean energy may also be 10% or so less than the expected value.

(*c*) *Conditions required for extraction of the beam.*

Little need be said about the use of the accelerated beam to bombard internal targets. Many forms of air-lock for introducing these without disturbing the main vacuum have been described, and the main problem—as in the design of anti-cathodes for high power x-ray tubes—is the dissipation of heat. In the M.I.T. cyclotron, for example, 1 ma. of ions at 12 MeV. has been obtained with a beam cross section of 0·2 cm², or 60 kw. per cm². A rapidly rotating tilted target which took over half the available current is shown in Figure 10 (*a*).

In Figure 10 (*b*) is shown a representative extraction system as designed for the Birmingham University cyclotron (so far untried). Particles entering the space between the tungsten septum and the deflector encounter a radial electrostatic field deflecting them outwards. If the field is F volts/cm., and the velocity of the ion is v, this will have the same effect as a reduction in the magnetic field by an amount $\Delta H = Fc/300v$. For most existing cyclotrons c/v will lie between 6 and 12 (corresponding to the range of 28 MeV. to 7 MeV. for deuterons). Then if we wish to increase the radius of curvature of the path of the particles by 10%, giving for most practical cases $\Delta H \simeq 1{,}500$ gauss, we shall need F between 40 and 80 thousand volts per cm., and an increase of even 10% may often be too small in a large cyclotron to take the particle out into a sufficiently weak field for it to escape.

The problem would not arise if we could use the magnetic field to the fullest extent. This would mean that the particles were accelerated until they reached the radius where Hr was a maximum. At this point $\partial H/\partial r = -H/r$, and beyond it H is falling off faster than r increases, so that a particle passing this point will as it were fly off the edge without external assistance. Close to $(Hr)_{max}$ a very small deflecting field would extract the particles and ensure that all came off at the

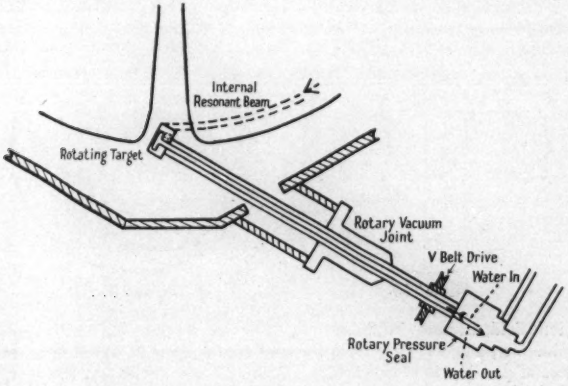


Figure 10 (*a*).

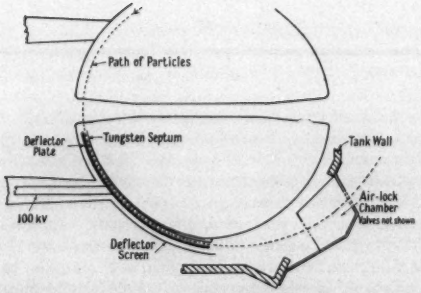


Figure 10 (*b*). Deflector system of Birmingham University cyclotron.

same azimuth and were not thrown out all around the tank. With even the most drastic shimming, however, it is impossible to get the curvature of magnetic field so great that we can reach $(Hr)_{max}$ before the value of H has fallen off by the 1% or less that is permissible within the phase limitations laid down in the last section. Hence deflection must start at a significantly smaller radius; in the Birmingham

cyclotron for example the particles must be dragged out 2½ cm. in radius before they are safely beyond the risk of recapture by the field.

As, for short lengths of deflector plate, the total radial displacement of the particle is proportional to the deflecting electrostatic field and to the square of the deflector length, it pays to have the longest practicable deflector. This, unfortunately, is not as long as one would like. As is shown in Figure 10 (*b*), for a cyclotron of any size the deflector is placed entirely within one of the dees. The voltage supply for it is brought up the centre of the dee stem. This is inconvenient but unavoidable, for only thus can the deflector be screened from the excessive R.F. voltage which would otherwise be induced upon it. Further, the gap between septum and deflector cannot be made less than 1 cm. or so owing to the spread of ion paths. As was pointed out in the last section, these paths are likely to be precessing with appreciable amplitude by the time they reach the exterior. Though this is most helpful in increasing the spacing between successive orbits, and hence reducing the proportion of particles intercepted by the leading edge of the septum, it leads to a spread in direction and energy which leads to a large amount of current striking either septum or deflector later on, unless the space between them is several times as large as the simple theory would seem to require. Hence we are committed to a deflector voltage which in a 60-inch cyclotron may reach 100 kv. or even more for the greatest efficiency of ion extraction.

The practical details of construction of septum and of deflector to cope with the great power densities encountered have little of special interest. It is normal practice to make the main part of the septum of tungsten sheet ground down at the leading edge, while the deflector is usually of water-cooled copper. Each may be bent to the optimum form as a result of empirical experiments on the regions of each which are most heavily bombarded. It may be useful to split the deflector into two parts, one above and one below the beam, a few inches round from the leading edge. Then the ions which are most deflected will get into the weak field between these two sections and will not be deflected further. This has recently been applied successfully in a small high-current cyclotron (Lofgren 1949).

(iii) *The Magnet*

The desirable form of the magnetic field has already been discussed. Consideration of costs of the magnet itself and of associated equipment shows that it is best to design a magnet for the highest field that can be used without saturating the steel of the poles. For a given particle energy therefore both mean magnetic field and final orbit radius are roughly determined, the field being greater and the radius less the narrower the magnet gap.

In Figure 11 are shown approximate profiles of two magnets of different gap sizes to give identical particle energies for a similar degree of saturation. The greater economy of the narrower gap is very apparent. Unfortunately the cost of the R.F. system varies with gap size in the opposite direction, and hence there will be some optimum ratio of magnet gap to diameter. Meanwhile it may be noted that for a given magnet shape the particle energy E increases nearly as r^2 and the weight W (and cost) as r^3. For an average shape $W \simeq kE^{3/2}$, where $k = 1{\cdot}5$ tons per $(\text{MeV})^{3/2}$.

It is not desired, any more than in the previous sections, to describe in detail the various magnet constructions in use. Magnets cooled by air, water and oil

are all used. In damp climates such as ours the first is convenient as it avoids the risk of condensation of water on the cooled coils if these fall below the dew point. It takes up a great deal of valuable space, however, and makes screening much more difficult (Figure 12*). If water cooling is used, air conditioning is essential, and this is expensive—although valuable in other ways, for example in the reduction of dust difficulties when the system is taken to pieces. Most new magnets are being made with oil cooling, which is much less likely than water cooling to cause condensation troubles, but is liable to suffer from messy leakages. The cyclotron at Princeton has recently been burnt out as a result of such a leakage.

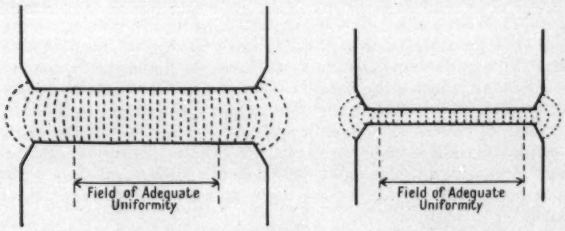


Figure 11.

A most important requirement is that the magnetic field should remain at the optimum value, the accuracy desirable being to the order of 1 part in 1,000. Control is usually maintained by means of the magnet-generator field-current. The main magnet-current or the actual magnetic field may be kept constant, or the useful output of the cyclotron may itself be used to keep the field at the optimum value. The latter method will enable the magnetic field to follow small changes in frequency, such as may occur in self-excited systems, or to compensate automatically for changes in R.F. voltage or other characteristics which have second-order effects on the value of magnetic field required. More complex amplifying equipment is needed, but systems using such methods are in use in Paris and elsewhere (Debraine and Simane 1948).

(iv) *The Radio-frequency System*

(*a*) *The generator.*

We saw in §2(ii *b*) that the R.F. voltage required on the dees increases very rapidly with the final particle energy required, owing to the need to complete the acceleration before the particles can get seriously out of phase with the R.F. field. In the small early cyclotrons, using a coil of copper tube as the inductance of the cyclotron resonant circuit, a simple self-excited push-pull oscillator of the tuned-grid tuned-plate type was used, with standard commercial valves.

As higher powers were required, a number of alternative systems came into use. A master-oscillator power-amplifier generator of the same general type as is used in broadcasting practice is used at Cambridge and some other laboratories. This is efficient and reliable but has the disadvantage that changes due to thermal

* For Figures 12, 14, 18 and 41 see Plates facing p. 320.

expansions etc. in the dee system can cause violent changes in the dee voltage and in the load-impedance presented to the generator.

The 'Q' of a large cyclotron is usually several thousand, so that a change of resonant frequency of 0·1% will have a disastrous effect on dee voltage, while if the generator frequency could follow this change only a small change in beam current would result. Some feedback can be used to control the frequency of the master oscillator, but the whole system then becomes rather complex.

The very low-loss circuits of large cyclotrons make the occurrence of high-frequency parasitic oscillations very difficult to stop where self-excited oscillators with any complexity of circuits are used, but the conventional self-excited oscillator using an extremely high-Q grid-circuit and a variety of wave-traps has been used successfully. At M.I.T. a neutralized push-pull oscillator with feedback excitation from the load has been developed, giving well over 20 kw. into the dee system for 40 kw. input.

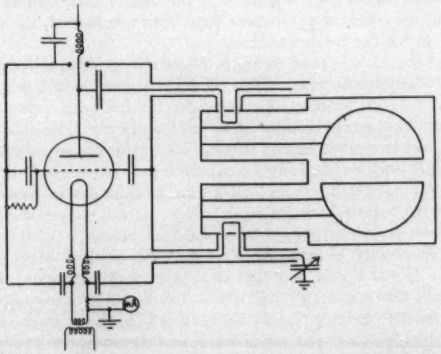


Figure 13. Radio-frequency circuit of Birmingham University cyclotron.

At Birmingham, where the 25 Mev. deuterons desired necessitate a considerably larger dee voltage than has so far been employed elsewhere, the generator takes the form of a single demountable grounded-grid oscillator capable of handling 300 kw. plate dissipation and using the dee system itself as the sole resonant circuit (see Figure 13). With this, over 60% of the D.C. input power has been fed into the cyclotron, the dee voltage being limited by electrical breakdown in the tank rather than by the supply. In this, as in other self-oscillator systems using a feed back from the dee system for excitation, help may be required in initiating oscillation unless the tank vacuum is extremely good. When the dee voltages reach a few hundred volts a 'Philips gauge' type of discharge may start. Electrons released from the tank cover plates oscillate up and down between the dees along the magnetic lines of force and ionize very efficiently any residual gas that is present. The currents involved are only of the order of milliamperes or less, but as for, say, 500 volts between the dees a power input of only $\frac{1}{4}$ watt should be needed to overcome ohmic losses, an extra watt or so in the discharge will reduce the amplitude by a very large factor and hence reduce the feedback voltage to an

inadequate level. The gas-discharge current varies only slowly with voltage, so that the power absorbed varies more or less linearly with dee voltage; hence as voltage increases the *proportion* of power taken becomes negligible. At large voltages the ionizing efficiency of the electrons falls off and the discharge current is actually reduced. Hence once a critical range of voltage is passed no further trouble occurs. The net effect apparent is that if oscillation stops for any reason it will not start again. The problem is solved by the use of a separate 'booster' oscillator with internal feedback and with an output of about a kilowatt, which can be used to drive the system through the danger region. This can be kept on continuously during the 'running in' of the cyclotron if precautions are taken to prevent excessive feedback from the main oscillator.

(*b*) *Construction of dees and resonant lines.*

In this section we shall consider the best design of the dees and lines for a given depth of vacuum tank. This is defined by the magnet size, and we shall discuss later, when all the essential parameters have been introduced, the way in which the best size of this can be determined.

It is clear from §2(ii*b*) that, owing to phase changes due to focusing requirements and to relativity effects, a critical value of the energy-gain per gap-transit is required before any particles at all can receive the full acceleration. To obtain a useful current this energy-gain must be exceeded by a comfortable margin.

We have thus two main factors to consider in the design of the dee system. The dee aperture must be large enough to permit the vertical oscillations discussed in §2(ii*b*) to occur without serious loss of ions, and the dee voltage must be high enough to give the required energy-gain per gap. It will be seen at once that these requirements are contradictory; a large aperture means a small clearance, and hence high capacitance beween the dees and the earthed cover-plates of the vacuum tank. It will also be seen that there is an optimum value; dees either of zero depth or of depth equal to that of the tank would give no output.

In a large fixed-frequency cyclotron a fall of $\frac{1}{2}\%$ in the magnetic field between the radius at which ions are in exact resonance and the exit slit will cause less phase-shift difficulty than will the relativity effect, so it is certainly profitable to use such a radial decrease to compress the beam vertically. We can then shape the dee aperture to follow the expected beam profile (see Figure 7) and thus save perhaps a third of the dee-to-earth capacitance with only a very moderate increase in the voltage required.

We shall now consider the relation between the peak dee-to-earth voltage V achieved and the required total power input P. If the resonant shunt-impedance of either dee to earth is R, the power needed for each dee is

$$P/2 = V^2/2R, \qquad \text{whence } V = \sqrt{(PR)}.$$

Hence our optimum design will be one giving a large value of R. It is frequently stated that a large 'Q' is the essential need of a cyclotron system. This is not the case; it is the high impedance that is required, not the high Q. The misconception appears because the measures taken to produce a large R (large copper conductors etc.) necessarily produce a large Q. The latter is in itself a nuisance; if the cyclotron were less sharply tuned it would be easier to drive. The distinction is of some, though not perhaps of great, practical importance, because the design for maximum R is different from that for maximum Q. A resonant coaxial line for

example of given outer diameter has a maximum R for a radius ratio of 9·1 : 1. The impedance here is 35% higher than that at the radius ratio 3·6 : 1, which gives maximum Q.

We can see from the relation $V = \sqrt{(PR)}$ why it is desirable to have two dees when power input is of importance. With the two dees taken for granted so far, the maximum equivalent voltage gain per gap is $2V = 2\sqrt{(PR)}$. If we had a single dee of resonant impedance R and total power input P', we should obtain a dee voltage of V' given by $P' = V'^2/2R$, whence $V' = \sqrt{(2P'R)}$, so that to get the same energy-gain per gap we should need twice as large a power input as with the two dees. The fact that we should also have twice as large a dee-to-earth voltage would also be serious in a big machine owing to spark-over problems. Discharges cannot jump from dee to dee because for their maintenance electrons would have to move across the magnetic field, which they are quite unable to do.

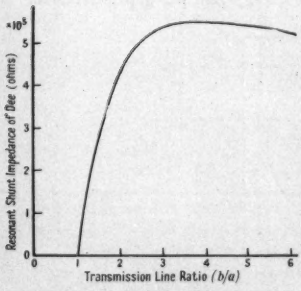


Figure 15 (*a*). Theoretical variation of dee resonant-impedance in Birmingham University cyclotron with variation of supporting transmission line.

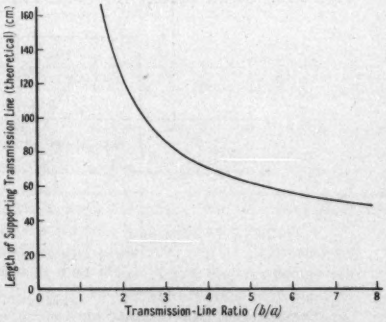


Figure 15 (*b*). Length of dee lines (for proton cyclotron shown in Figure 19).

Now let us return to the question of obtaining an optimum dee impedance. For the frequencies needed for accelerating deuterons or α-particles, the dee diameter will be short compared with the wavelength (20–30 metres). The main power-loss in the system will therefore occur in the resonant lines used. In most existing large fixed-frequency machines each of these lines consists of a pair of coaxial copper tubes (see Figure 14). The losses in a shielded pair would be 5–10% less, but the mechanical construction may be less convenient. For each line the problem is to produce an inductive reactance to match the capacitive reactance of the dee at the resonant frequency with the minimum equivalent series resistance. Since the losses in the line vary inversely with the outer diameter, the most economical line will be the largest that the space available will permit. The best diameter for the inner depends on the dee reactance. As indicated above, for a simple quarter-wave line, corresponding to the case of zero dee capacitance, the inner diameter in the coaxial case would be 1/9·1 times the outer diameter. Since the dee has in fact a large capacitance, the optimum ratio is closer to unity owing to the increased current which the inner has to carry and to its reduced

length. In Figure 15 (*a*) is shown how dee-impedance changes with the diameter-ratio of the transmission line for dees of the form in use in Birmingham. The optimum is extremely flat; it will be seen from the figure that with the existing ratio of 3:1 the impedance is only 3% below the maximum obtainable with a 4:1 (optimum) ratio. For a 5% drop of impedance the ratio limits are 2·8:1 and 5·8:1. For practical reasons, such as the need for mechanical rigidity and to be able to lead in the H.T. line for the deflector plate, it is usual to use an inner line diameter appreciably larger than the optimum, but within the 5% limit. The increased length required, owing to the lower inductance per unit length, is usually unimportant and may even be more convenient (see Figure 15 (*b*)).

The shape of the dees seen in the plan is next to be considered. It has for some time been usual to 'taper off' the dees a little to reduce dee-to-earth capacitance, as is noticeable in Figure 14. This process could, it is believed with advantage, be carried much further. This would lead to dees of the shape indicated in Figure 16. If the process is taken too far the energy-gain per gap is reduced

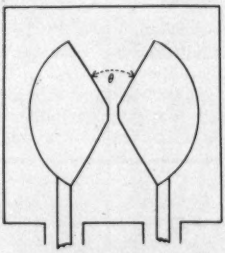


Figure 16.

below the normal value of $2V\cos\phi$, where ϕ is the phase of transit at the centre of the gap, to $2V\beta(\theta)\cos\phi$, where $\beta(\theta)$ represents the effect of field change during the time of transit of the now much enlarged gap. If the depths of dee and tank are small, $\beta(\theta)=\cos\theta/2$; for practical systems it will be slightly larger. It will be seen that when θ is small the change in β varies as θ^2 while the change in capacitance varies nearly as θ. For a given power input therefore the gain due to decreased capacitance at first predominates over the loss due to the reduction in the factor $\beta(\theta)$. In Figure 17 is shown the way in which the power input required varies with θ for a given energy-gain per revolution. The arrow marks a point equivalent in capacitance to the shape of the gap at present in actual use in Birmingham or in the 60-inch machine at Berkeley. It will be seen that over 15% more power is required than would be required by the optimum design. The real advantage, however, of the cut-back dees would be that a good space free of R.F. field would be left for insertion of targets etc. A minor advantage is that with radial dee-edges there would be no radial component of R.F. field as there is with parallel edges, though it is unlikely that this is large enough in any case to contribute seriously to any undesirable radial oscillations.

A practical disadvantage in the case of machines driven by self-oscillators is that the dee-to-dee capacitance is so far reduced that very slight unbalance between

dees may lead to very different amplitudes of oscillation. This wastes power and may make the beam extraction extremely difficult, but could be dealt with by introducing extra coupling between the dee lines elsewhere in the system.

It is hardly necessary to say that all interior surfaces should be copper and all connections should be extremely good. Currents of the order of 10,000 amps flow in the lines of a 60-inch cyclotron and the lines of current flow are entirely determined by the relative reactances of different paths. Hence if there are good

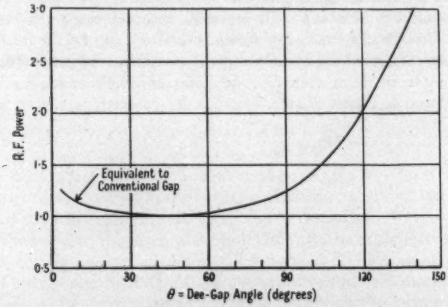


Figure 17. R.F. power required for fixed energy-transfer per gap-transit.

and bad regions in parallel along a line of contact the current will flow over good and bad alike and dissipate excessive power in the high-resistance regions.

The complicated nature of water-cooling requirements can be illustrated by Figure 18. If the cooling fails on any one area of dee or line the area concerned will overheat, probably enough to melt, so a flow-operated relay is employed on every separate water circuit to prevent power from being applied at the wrong time.

(v) *Limits of the Fixed-frequency Cyclotron*

We have now the necessary data to discuss the limiting energies practically obtainable. The problem can be well illustrated by considering the re-design of the Birmingham 60-inch cyclotron as a proton accelerator.

Suppose first that the existing tank depth (23 cm. inside the cooled-copper liners) is taken, with the maximum usable magnetic field (18,000 gauss). The mean resonance-wavelength will be 11 metres. To keep reasonably uniform voltages along the dee edges the dees must be centre-fed, but this is not inconvenient as the resonant lines will be short enough to lie inside the yoke of the magnet, leaving both sides free. In Figure 19 is shown a drawing of a proposed cyclotron, with dimensions optimized according to the principles discussed in the last section. Measurements by Dr. J. Walker on a 1/20 scale model showed resonant shunt-impedance in satisfactory agreement with the calculated values. With this structure, and using the improbably high power input of 200 kw., the energy gain per gap by a proton would be almost 700 kev.

If we modify the formula given by Rose (1938) to take account of the peculiarities of the initial paths discussed in § 2 (ii *a*), we find a maximum obtainable energy of 41 Mev. The mass of the protons will then be $4\frac{1}{4}\%$ greater than the rest mass.

The diameter of the final path will, however, be only 105 cm. (41 in.), while the magnetic field could be made sufficiently uniform out to perhaps 136 cm. (54 in.). We are thus using only a fraction of the magnet capacity; a magnet of half the weight and cost would give the necessary field out to 105 cm. diameter. On the other hand, even if voltage breakdown did not occur (and the dee-voltage already assumed might well be impracticable for this reason) a power of *more* than 200 kw. could hardly be considered. To use the full 136 cm. diameter would require a continuous-wave radio-frequency power of about 10 megawatts.

An advantage can be gained, for work on internal targets, by widening the magnet gap, though the resulting slow fall-off of magnetic field near $(Hr)_{max}$ would make extraction of the beam almost impossible. If the usable tank depth were pushed up from 23 cm. to 45 cm. the diameter of field of adequate uniformity would be reduced to about 122 cm. and the dee-to-earth capacitance very much

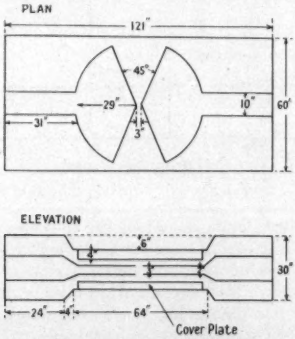


Figure 19. Proposed design for proton cyclotron to fit Birmingham University 60-inch magnet.

reduced. This would give about 50 MeV. protons at about 118 cm., but the R.F. gap voltage would have increased to over 1 million volts, and with the increased discharge space available this could certainly not be maintained with existing techniques. Hence, although for the present magnet the optimum tank depth is a little over 45 cm., the magnet is being used most uneconomically.

It must be remarked that we have been discussing only the most-favoured proton, and the great difficulties inherent in running a large fixed-frequency cyclotron make it quite undesirable to use such a machine unless large currents are attainable. For this purpose the useful upper limit for proton-acceleration is likely to be below 35 MeV., for which a magnet pole of little over 40 in. diameter would be sufficient at high fields. Even here beam extraction might be very difficult.

For deuterons a higher energy beam might be obtained owing to the later onset of relativity effects, although the lower-frequency lines will be bulkier and less conveniently spaced. The practical difficulties in handling large R.F. powers are, however, very great, and it is to be noted that the beam has not yet been extracted from the Birmingham machine (25 MeV. deuterons) four years after work on the

main part of the construction was started. Many non-essential difficulties, such as a cast bronze 'vacuum' tank (nine months required to achieve $\frac{1}{2} \times 10^{-5}$ mm. pressure) and a shunt impedance of half the value which could be obtained in a new design, have contributed; but it is clear that even in this region

$$\frac{d(\text{human effort})}{d(\text{particle energy})}$$

is very large, and it may be decided to reduce the magnetic field and frequency so as to lower the final energy to 20 MEV. It may well prove that the 20 MEV. deuterons obtainable from the largest steadily operating cyclotrons, at Berkeley or the Carnegie Institution, represent the economical limits for fixed-frequency machines.

§3. THE FREQUENCY-MODULATED CYCLOTRON, OR SYNCHRO-CYCLOTRON

We have seen in the last section that the angular velocity He/M of accelerated particles in a cyclotron falls off as the energy increases, as a result both of relativistic effects and of the inescapable need for magnetic focusing. It was realized very early that if the angular frequency of the accelerating electric field could be reduced progressively by the appropriate amount during the acceleration of a particle, the limitation of energy obtainable from the cyclotron would be removed. For some years, however, no practical attempt was made to vary the frequency of a cyclotron during operation. This was due to the fact that it was believed that the variation of frequency would have to be such as to maintain a constant phase of rotation for particles receiving nearly the maximum energy at each gap transit.

For existing machines this would have meant a modulation frequency in the region of 100 kc/s. and having a far-from-sinusoidal wave-form. With a large cyclotron, electronic methods of frequency variation would have required devices with many hundreds of amperes of electron emission, and reactances mechanically variable at such a frequency would present even greater difficulties. Lowering the accelerating voltages would simplify matters by increasing the acceleration time, but it was believed that the increased number of revolutions would make it necessary to keep the frequency correct at any instant of the acceleration to an impossible accuracy.

The solution to the difficulty was seen independently by workers in at least three places, the first two being Veksler (1944) and McMillan (1945). They showed that acceleration of particles could be carried out successfully by progressively reducing the frequency of the electric field, however slowly this reduction took place, so long as the total reduction was sufficiently great. This is explained in the next section.

(i) *Effect of Frequency Variation*

Consider a particle starting from the centre of a variable-frequency cyclotron at the instant when the angular frequency ω_E of the radio-frequency electric field is $H_0 e/M_0$ so that slow-moving particles of rest mass M_0 are resonant in the central field H_0. Then if the angular frequency ω_E is falling only very slowly the initial paths of the particles will be practically as discussed in §2(ii *a*) (see Figure 5). The particles will be bunched together and will be rotating a little ahead in phase of the R.F. field, i.e. crossing the line of symmetry between the dees a little before

the R.F. field reaches its maximum. In a constant-frequency R.F. field the particles will gradually fall behind in phase owing to decrease of H and increase of M until after a definite maximum radius acceleration no longer occurs. The particles would continue to fall behind in phase until they began to be decelerated and finally return to the centre. The variation of energy with time for two different dee-voltages during the process is shown in Figure 20.

Clearly, if during their initial period of acceleration the radio frequency is decreasing slowly, they will not fall behind in phase so fast and will reach a larger radius before they get 90° behind in phase and then begin to be decelerated. Even more important than this, however, is the fact that, when decelerated, and therefore returning to a larger magnetic field and lower mass, they begin to catch up in phase again much faster than they fell back on the way out, since the R.F. is still decreasing while the resonant frequency He/M is increasing. Consequently they get 90° out of phase with the decelerating field again before they reach the centre and while they still have considerable kinetic energy. They are now accelerated again and, since the change of R.F. and of He/M are in the same direction, they remain in the half cycle of phase-range for acceleration for longer than in the

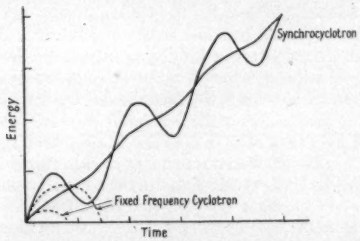


Figure 20. Variation of particle energy with time for two dee voltages.

preceding decelerating half cycle. They therefore reach a larger radius than before, after which the process occurs again, and continues to do so until the particles reach the maximum radius available. The variation of energy with time is again shown in Figure 20.

The particle motion then, instead of being a continuously increasing spiral, follows an orbit whose radius oscillates steadily about a continuously expanding 'stable' orbit defined by the fact that the value of He/M characterizing the orbit must equal the instantaneous value of ω_E.

The larger the R.F. voltage and the slower the change in ω_E, the more violent are the oscillations. As the R.F. voltage is reduced, the oscillations get smaller until we reach the state where the particles can only just gain energy fast enough for their frequency of revolution to slow down at the same rate as ω_E. The particles will then remain nearly at the optimum phase throughout the whole acceleration. The final energy reached and the time taken to reach it will both be almost unchanged, the first being always the energy corresponding to $(Hr)_{\max}$ and the second the time it takes for ω_E to fall from H_0e/M_0 to $(He/M)_{\max}$ (see Figure 20).

As far as the main acceleration is concerned, therefore, the R.F. voltage only needs to be above a critical minimum, which can be reduced as much as desired by

reducing $\partial\omega_E/\partial t$. If ω_E is changed mechanically, as is usually done, it is difficult to reduce the total acceleration time below about one millisecond. Then for an accelerating frequency in the region of 10 Mc/s. we have some 10,000 revolutions in which the final energy may be obtained and the energy required per revolution is about a hundred times smaller than in a normal fixed-frequency case. The practical effects of this will be discussed in the next section.

(ii) *Early Experimental Work*

When work began on the 184-inch cyclotron at Berkeley before the war it was intended to run it as a fixed-frequency machine. It was hoped that pulsed voltages of the order of one million could be obtained on the dees and that deuterons could be accelerated to energies of the order of 100 MeV. This would have been very difficult, and when work on it could be begun again after the war McMillan suggested the use of frequency modulation, relying on the 'phase-stability' principle outlined above. In order to test this proposal the Berkeley 37-inch cyclotron was converted for frequency modulation (Richardson *et al.* 1948). The magnetic field was adjusted so as to require a 13% decrease in dee frequency between ion source and exit radius. This change gave the same change in ion angular velocity as would be produced in the 184-inch machine by the relativistic increase of mass. It also gave an exceedingly high degree of magnetic focusing

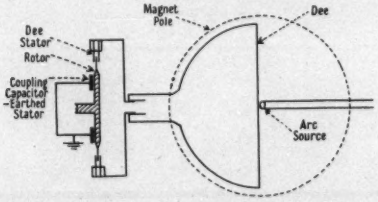


Figure 21. 37-inch cyclotron dee and capacitor for tests of frequency modulation. Vacuum enclosure not shown.

which would not be present in the larger cyclotron, but there was no way of avoiding this. The difference was not expected to be, and was not in fact, of great importance.

A single dee only was used. The much lower voltages needed by the synchro-cyclotron made the consequent doubling of power input quite unimportant compared with the resulting simplification of the problem of modulating the frequency.

In Figure 21 is shown a plan view of the system. The frequency modulation was produced by a rotating capacitor connected to the end of a half-wave resonant line which formed the dee stem (Schmidt 1946). The shape of the frequency–time curve could be adjusted during operation by moving part of the stator of the capacitor.

The ion source could be pulsed, the pulse width being variable from 3 to 50 μsec., and the start of the pulse being adjustable to any desired instant of the frequency-modulated cycle.

The use of an open-ended half-wave line as the dee resonant circuit made it easy to apply a positive D.C. bias to the dee of 1,000 to 2,000 volts. This was very helpful in preventing gas discharges by clearing up residual ions. It also increased the beam output by a factor of ten with 3 kv. of radio frequency, falling to a factor of 1·2 at 10 kv. This may be due to the fact that at the lower voltage a serious degree of space-charge limitation exists. The first 180° of the orbit after leaving the ion source will be increased appreciably in size by the superimposed D.C. field for the ions starting off in the positive half-cycle of dee potential. This would increase materially the space-charge-limited current as well as getting these ions rather more rapidly through the dangerous part of their paths, where the electric defocusing forces have a serious influence.

The application of frequency modulation was immediately successful. Particles could be accelerated to the full energy (7·5 MEV. for deuterons and 14·6 MEV. for protons) over the whole range of modulation frequency which was tried (Figure 22). The ion stability was so great that a few particles could even be accelerated if the capacitor was rotated by hand to an intermediate position, left there for a brief interval, and then rotated the rest of the way. In this case ions must have revolved millions of times in paths oscillating about a stationary mean orbit before being taken up to their final energy.

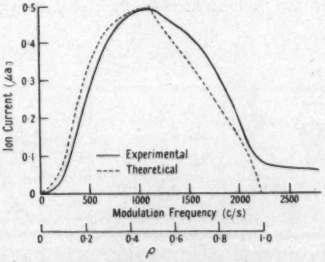


Figure 22. Variation of ion current with modulation frequency in 37-inch synchro-cyclotron.

An interesting point is that some ions received full energy even when the frequency changed so fast that it was impossible for them to do so in one modulation cycle (see Figure 22). It was shown, both theoretically and experimentally, that when ions were 'left behind' by the too-rapidly falling frequency they rotated without serious gain or loss of energy until the dee frequency again approached their rotation frequency. By this time their phases had become uniformly distributed so that about half were in suitable phase to be further accelerated. If this process were repeated N times before the ions reached the target, the current would be reduced to about $1/2^N$ of its value. This agrees very fairly with the experimental results.

When the work was started it was believed that, though an internal circulating ion beam should be obtainable, a conventional D.C. deflector would be useless as, for 7 MEV. deuterons for example, with a maximum energy increase per revolution of 4 kev., the change of radius Δr for successive turns of the ion path would be about 0·12 mm. It was found experimentally, however, by examining the

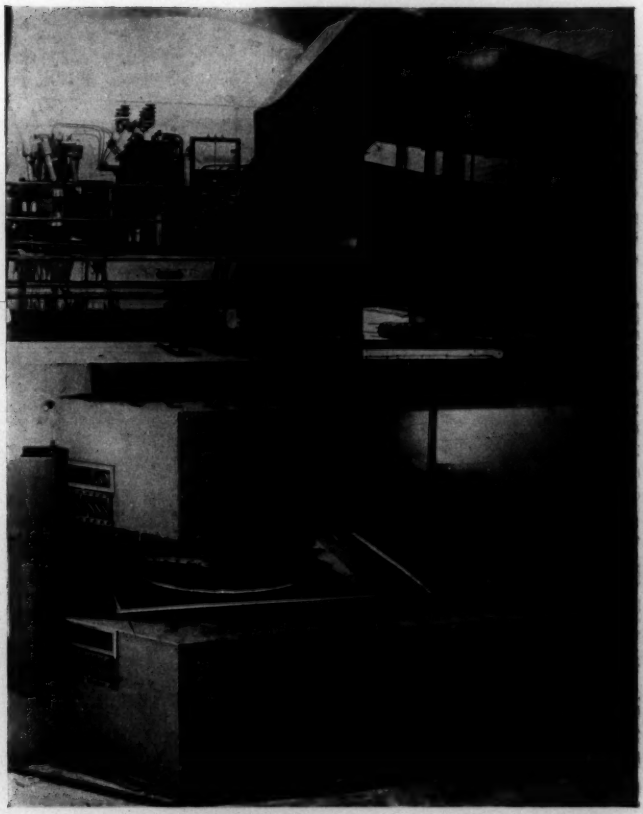


Figure 12. Photograph of Birmingham University magnet. (Photo : J. W. Blamey.)

To face page 320

Figure 14. Photograph of dees and transmission lines in Birmingham University cyclotron (Photo : Dr. D. M. Millest.)

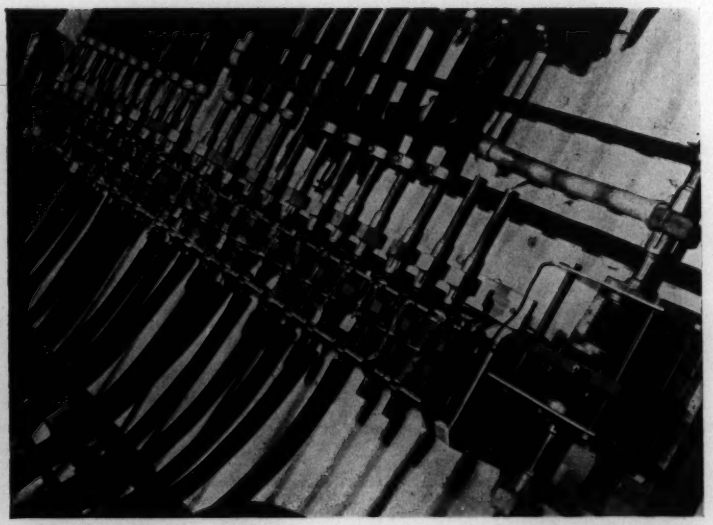


Figure 18. Photograph of some of the water-flow-operated switches on the water-cooling lines of the Birmingham University cyclotron. (Photo: R. S. Johnson.)

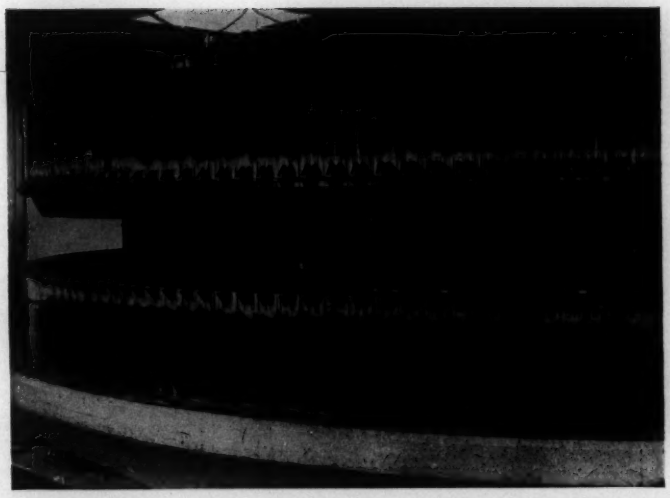


Figure 41. Birmingham University proton synchrotron. The magnet yoke is shown with the windings and one pole piece complete and with part of the other pole and a section of the vacuum chamber with pumping ports.

distribution of β-activity for a copper target, that the increase in distance from the centre on successive turns could be greater than 3 mm. This is due to the precession of orbits mentioned in §§2 (ii), (*b*) and (*c*), and made it possible to extract up to 10% of the deuteron beam with an ordinary D.C. deflector, as against perhaps 25% possible (though not very often achieved) in a fixed-frequency cyclotron.

The practicability of applying frequency modulation to cyclotrons was thus demonstrated, and in the next section we shall describe some of the results that have been achieved.

(iii) *The present Position of the Synchro-cyclotron*

The 184-inch machine at Berkeley was the first, and is still the largest, of existing synchro-cyclotrons. It is therefore our main source of knowledge of their operational characteristics. A diagram of the machine is shown in Figure 24. For deuteron acceleration the following represent typical operating conditions. The single dee was run at about 20 kv. peak, with a central magnetic field of 15,000 gauss, falling by about 5% at 81 in. radius. The modulation frequency was 100 c/s., obtained by rotating a 24-toothed capacitor at 250 r.p.m. and giving a frequency range of 12·6 Mc/s. to 9·0 Mc/s., more than twice that required. Then with an arc source pulsed for 100 μsec. at the optimum point of the frequency-modulation cycle, a mean current of 1·0 μa. of 190 MeV. deuterons was obtained at 81 in. radius. Although this current seems small, it was only possible to run the machine at full output for 5 minutes per shift with the original screening owing to the large flux of very high energy neutrons.

Serious trouble due to gas discharges was encountered at first owing to the very large volume of the machine. (Similarity laws would suggest that the pressure should be varied inversely as the tank depth for constant ionization density, and the reduction in pressure required to satisfy this condition for so large a machine was not obtainable, 10^{-5} mm. being the practical minimum.) The trouble was controlled by applying 1 kv. D.C. of negative bias to the dee. The average energy gain per revolution was found to be about half of the peak possible energy gain, in good agreement with the maximum of the curves shown in Figure 22. Hence the particles revolved about 10,000 times during acceleration.

A number of interesting experimental results were obtained concerning the behaviour of the beam. Measurement of the 'acceptance time' during which particles leaving the source might be accelerated to full energy showed that the particles could be accepted from the source for up to twenty or so times *longer* than theory suggested. This was demonstrated by releasing a very short pulse of ions at a variable phase of the modulation cycle, and means that some particles were fully accelerated which were released at times when the frequency was much higher than the correct value for low-velocity ions. This can be explained on the assumption that such ions are at first accelerated and then, as they get out of phase, returned to near the ion source in time to be picked up by the accelerating field when it reaches the correct frequency. This process might happen several times, resulting in an extremely long acceptance time. This is analogous to the behaviour which results in successful acceleration even when the frequency is changed too rapidly, as described in the previous section.

When the variation of target current with time was displayed on an oscillograph it was found that, instead of being received in one pulse corresponding to the arc-source pulse for each modulation cycle, it appeared in a series of a dozen

or more pulses 10–12 μsec. apart. The *number* of pulses increased with radius, showing that they were not produced at the source. This phenomenon appears to be due to the fact that the phase of the precession discussed in § 2 (ii *c*) is roughly the same for all ions, the calculated precessional period being found to be about 10 μsec., in good agreement with the observed results.

The current remained nearly constant between 20 in. and 80 in. radius. Unfortunately it was then found to fall off rapidly, and dropped completely to zero at 82 in., although the value of Hr did not reach a maximum till 85 in. radius. This is the point where $-(r/H)(dH/dr)$ (n say) reaches 1. At 82 in. n, from magnetic field measurements, was only a little over 0·2. The explanation of this appears to lie in the fact that the very large magnetic restoring force which exists when the field H is changing so fast increases the frequency of vertical oscillation until, when $n = 0{\cdot}2$, it becomes equal to exactly half the frequency of the radial oscillations, corresponding to the precession mentioned above. The energies of the radial oscillations are large, amplitudes of up to 5 in. being observed, and although the coupling between the two oscillations is not very close at resonance, enough energy may be fed into the vertical ones to cause most of the ions to strike the dee.

Proof of the loss of ions to the dees in this region was obtained by allowing the beam to strike a probe of the shape shown in Figure 23. The distribution of radioactivity for three different radii is shown.

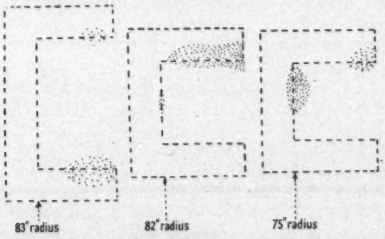


Figure 23. Distribution of radioactivity on copper target probes in 184-inch synchro-cyclotron.

The disappearance of the particles at $n = 0{\cdot}2$ complicates very severely the extraction of the beam. If the beam persisted to near $n = 1$ at 85 in. a very small deflector-voltage would take it 'over the edge' of the magnetic field at any suitable part of the circumference. As it is, a deflecting field of some 100,000 v/cm. over a length of about 14 ft. is needed. To avoid the great loss of ions produced if a conventional septum and D.C. deflector is employed, a pulsed deflector of the form shown in Figure 24 was designed. The effective capacitance between the two halves of this was about 270 $\mu\mu$F., and it was successfully pulsed to 300 kv. with a rise-time of 0·1 μsec. followed by a duration of about 0·2 μsec. If the deflector pulled the ions outward in the ordinary way they would have emerged in the region of the rotating capacitor. This would have been most inconvenient, and hence the ions were deflected *inwards*. The resultant displacement of their centre of rotation was sufficient to cause them to increase their radius by some

2 in. at about 350° round from the leading end of the deflector system, when they entered a magnetic shield which reduced H by about 70% and allowed the ions to emerge through an unimpeded section of the tank wall.

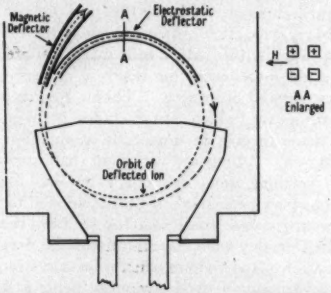


Figure 24. 184-inch synchro-cyclotron showing deflector system.

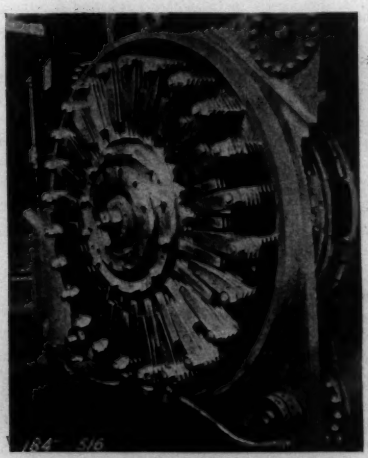


Figure 25. Photograph of rotating capacitor used on the 184-inch synchro-cyclotron when accelerating deuterons or α-particles.
(*By courtesy of F. H. Schmidt* (*MacKenzie, Schmidt* et al., 1949).)

As in the case of the 37-inch cyclotron, power was obtained from a grounded-grid oscillator. At 20 kv. dee-amplitude the R.F. power delivered was 20 kw. at 60% efficiency, the oscillator being capable of considerably greater output (Burkig, Hubbard and MacKenzie 1949). As before, the rotary condenser (see Figure 25)

was placed at the end of a 'half-wave line', the toothed stator being joined to the dee line and the rotor connected capacitively to earth.

After about two years of operation the 184-inch cyclotron has been changed over to operate at a higher frequency in order to accelerate protons. The practical problems are very much more severe, as for this purpose a minimum frequency-range of 22·9 to 15·8 Mc/s. is required. An ingenious system devised by MacKenzie (1948) has succeeded in doing this with a wide margin of safety. The capacitor is earthed through a low-impedance line nearly a quarter-wavelength long, at the highest frequency, instead of directly. Then a frequency range of 2·4 to 1 can be covered continuously (which just includes both proton and deuteron ranges) and can be made to give an amplitude modulation of only 20% over whichever range is chosen. A long and very rigid shaft, carrying 72 rotor plates with 6 teeth each and weighing about two tons, is required to give the full capacitance change. Protons have been successfully accelerated to 350 MeV., no major difference in the operating characteristics having yet been recorded.

The success of the Berkeley work has stimulated the development of a large number of synchro-cyclotrons elsewhere, chiefly for acceleration of protons, the larger ones, whose construction is well advanced, being at Chicago (170-inch), Columbia (164-inch), Rochester (130-inch), Harvard (112-inch) and Harwell (110-inch). These vary considerably in the form of rotating capacitor. At Rochester for example a rotor made of copper-plated zircon ceramic is used so that the capacitor can be placed in the edge of the magnetic field without serious eddy-current losses. Hence the stator can be joined directly to the back of the dee, which is supported near each corner by a short transmission line. This machine is already working, and by February 1949 was giving a beam of 0·1 μa. of protons of approximately 240 MeV. (Barnes *et al.* 1949).

At Columbia the large single dee has a long row of toothed stators right along each leading edge with a shaft carrying a similar row of two-bladed rotors (about 50 blades per foot) between each edge and the earthed tank. By using a considerably shallower tank (about 1/3 the depth of the Berkeley machine) it is hoped that a proton energy considerably larger than 350 MeV. can be obtained from this in spite of the smaller magnet pole diameter. In view of the fact that volume-reducing screens have had to be introduced in the Berkeley tank to prevent discharges, and the fact that high voltages and high powers are no longer required, this seems a very desirable development which is likely to become general.

The latest machine to work is that at Harwell (Pickavance, Adams and Snowden 1950), which has produced about 0·5 μa. of protons at 180 MeV., and a number of further large machines both in Europe and America should be operating within a year or two. The largest British synchro-cyclotron proposed is the 140-inch machine at Liverpool, which, like the larger American ones, should be able to give a considerable output of π-mesons, which the Harwell cyclotron will only just be able to produce, if at all.

§4. BETATRONS

(i) *Introduction*

It is now generally recognized that betatrons will not give such high ultimate particle energies as can electron synchrotrons. They are, however, capable of a much higher current output—though at higher cost—for energies within their range, and they are of considerable interest in any general discussion on cyclic

accelerators. Not only are they of great historical importance in paving the way for the more economical synchrotron and in solving the outstanding problems associated with all these types of machines, but nearly all electron synchrotrons use the betatron principle in the initial stages of acceleration. The problems of injecting particles into a synchrotron are identical with those of a betatron, and it is the injection process which is the most troublesome in these devices.

The suggestion of a betatron was first put forward by Wideröe and Breit and Tuve in 1928 (Wideröe 1928, Breit and Tuve 1928). After that date several experimenters tried unsuccessfully to accelerate electrons by this principle. It was not until 1941 that Kerst (1941), working on the theory of Kerst and Serber (1941), successfully overcame the injection and stability problems. He accelerated electrons to 2 MeV. Since then other workers have followed his example, and there is quite a large number of betatrons now in use. A very good survey of the development of the betatron is given by Kerst (1946).

(ii) *Principle of Operation*

The components of a betatron are a magnet and power source, a vacuum system and an electron gun. The electrons are constrained to move in an annular space inside a doughnut-like vacuum chamber by means of a magnetic field. As they are accelerated the magnetic field rises in step with their momentum and retains them inside the annular space. The magnetic field also accelerates the particles. The major part of the total changing flux passes through the centre yoke as shown in Figure 26. In this way an electric field is induced around the

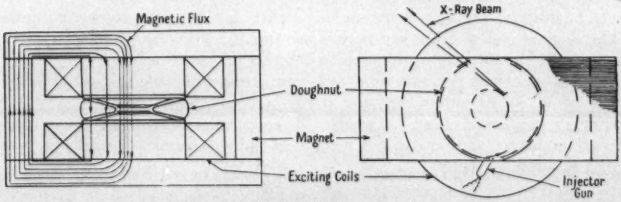


Figure 26. Principle of the betatron.

annulus by magnetic induction which continuously accelerates the electrons. Injection is achieved by directing a stream of electrons tangentially into the annular space either from the outside or the inside radius of the doughnut, but from inside the magnetic field.

There are clearly several general principles which have to be satisfied before such a machine could be expected to work.

In the first place, the induced accelerating electric field must be of just the correct strength at a given orbit (the stable orbit) to ensure that the particles remain on that orbit in an increasing magnetic field, i.e. the flux linking that orbit must be of the correct amount. There must also be some form of stability to the particle motion so that particles not on this orbit will not fly out of the doughnut but will oscillate about the stable orbit. These conditions were found to be satisfied by

Wideröe (1928), Walton (1929), and Kerst and Serber (1941) if the total flux linking the stable orbit is given by

$$\Phi = 2\pi R^2 H_0, \qquad \ldots\ldots(6)$$

where Φ is the linking flux, H_0 the flux density at the stable orbit and R the radius of the stable orbit. This condition holds only for a conventional betatron such as is used in the initial stages of an electron synchrotron. Later improvements in the betatron have made it unnecessary for this condition to be satisfied. A more general requirement is that the rate of change of flux through the centre orbit should be given by

$$\frac{\partial \Phi}{\partial t} = 2\pi R \frac{\partial H_0}{\partial t}. \qquad \ldots\ldots(7)$$

Secondly, it is clearly most important that the magnetic field should have focusing properties in both the horizontal and vertical directions. This is achieved if the vertical component of field along a radius has the form $H_z = H(R_0/R)^n$, where $0 < n < 1$ (Kerst and Serber 1941). Values of n usually chosen are between 0·5 and 0·75. In such a magnetic field, if an electron is deflected from a circular path it will undergo an oscillation about that circular path, in the same way as do the ions in a cyclotron as described in § 2 (ii *c*), the oscillation amplitude decreasing with increasing particle energy. For an oscillation in the horizontal plane the frequency of oscillation is given by $\nu(1-n)^{1/2}$, where ν is the frequency of revolution of the particle, and for an oscillation in the vertical plane the frequency is $\nu\sqrt{n}$. The amplitude of these oscillations varies as $H^{-1/2}$.

Finally, the problem arises as to how the electrons can be injected and be made to miss the injection apparatus on subsequent revolutions. This is a problem which, although partially overcome in practice, has not yet a complete solution. The original idea of Kerst and Serber was that the electrons, if injected circumferentially, should follow orbits which contract for two reasons:

(*a*) For a certain period of time the instantaneous orbits of the injected particles will be of larger radius than the stable or equilibrium orbit. The stability principle mentioned above will apply. The electrons on such orbits will receive less than the required energy per revolution and will spiral in towards the stable orbit at a rate given by

$$\frac{dx}{x} = \frac{1}{2}\frac{dE}{E}, \qquad \ldots\ldots(8)$$

where x is the displacement of the instantaneous orbit from the equilibrium orbit and dx is the shift of the circle toward the equilibrium circle, while the electron's energy increases by the fraction dE/E.

(*b*) Secondly, the oscillations induced during the injection process by particles entering the chamber at a small angle to a tangent are damped at a rate given by

$$\frac{da}{a} = -\frac{1}{4}\frac{dE}{E}, \qquad \ldots\ldots(9)$$

where da/a is the fractional decrease in amplitude of the oscillation about the instantaneous circle.

If this were the case, then it is to be expected that the lower the injection energy the greater would be the rate of contraction of the orbits etc., and the greater the current accepted. Experimentally, the reverse is true; in fact, the results of

injection experiments are most confusing. Increased yields have been obtained by Kerst and others with the use of contractor coils. These are coils which are situated above and below the equilibrium orbit. At the time of injection, current is pulsed in the same direction as the electron current injected from the outside. The field is increased temporarily, thus causing the electrons to spiral inwards. The success of this device has led Kerst (1948) to propose a theory of injection which has some attractive features. He argues that the electron stream loses energy by creating its own magnetic field, and thus causes itself to contract. For reasonable values of current etc. a contraction of approximately 4 mm. during a few revolutions is calculated to be possible. This is supported by the fact that the output of many betatrons increases more rapidly than linearly with the injection current.

Reports show that in many cases current is unaffected by a projection of several millimetres extending beyond the gun into the chamber. Biasing the conducting coating negatively near the electron gun increases the output in the Manchester machine (Heymann 1949). It is clear that the injection process is a very complicated one, and is not likely to be explained by a single simple hypothesis.

Finally it should be pointed out that since the acceleration is continuous the velocity–energy relation does not enter into the operation of a betatron, and the method is equally good for relativistic and non-relativistic energies.

(iii) *Physical Construction*

(*a*) *Magnet*. The magnet clearly plays by far the most important part in a betatron (and in a synchrotron). It is also the magnet which gives the greatest number of troubles. Most betatrons have a magnet whose general construction is as in Figure 26. The magnetic field is made to oscillate at the mains frequency (50 or 60 c/s., or sometimes as high as 600 c/s.) by making the magnet inductance resonate with a bank of capacitors. This means that thin laminations of transformer iron are used in the construction. A maximum flux density of about 5,000 gauss is usually used in the doughnut region. It is not possible to go much higher than this and yet satisfy the flux condition in a conventional betatron. This condition is that the flux enclosed by the stable orbit shall be the same as if the entire area inside the orbit were at twice the flux density of the field at the orbit. Since the characteristics of iron begin to deteriorate a little above 15,000 gauss there is a limit on the guiding field of between 5,000–7,500 gauss. It is clear that the conventional betatron requires a great deal of iron, mainly for the accelerating flux. Several schemes have been proposed for reducing the amount of iron required (Westendorp 1945, Kerst 1945, Amaldi and Ferretti 1946). If the guide field is biased negatively the centre flux can be made to pass from its maximum value in one direction to its maximum value in the other direction in a half cycle. Thus the rate of change of field enclosing the orbit is the same as before and consequently the maximum field required in principle inside the orbit is now only the same as at the orbit. In this way a saving of 50% of the iron required is achieved. Special precautions such as flux forcing (Kerst 1945) are required to keep the rate of change of flux in step with the guiding field.

The greatest difficulty encountered in the construction of betatrons and electron synchrotrons is that of obtaining a sufficiently uniform magnetic field at injection when the field strengths are only of the order of 10 gauss. It is clear that in order to retain the electrons inside the doughnut the magnetic field must

be sufficiently uniform circumferentially as inhomogeneities cause oscillations of the type mentioned above. It might be of some interest to summarize the sort of errors which may occur. These are:

(1) Variable remanence from point to point. This can be very serious if the magnet is pulsed as in large synchrotrons, as the remanent field can vary from place to place by as much as a factor of two, and its mean value may be as high as the injection field itself. The remedy for this sort of trouble is to place coils over the pole tips in the regions where the effect is bad or else to wind coils around the various segment yokes if the betatron is constructed in this way. Currents are then made to flow in these coils so as to correct the remanent effect at injection time. Sometimes currents are simply induced in these coils and their amplitudes suitably controlled. It is clearly important that the magnet always follows the same magnetic cycle.

(2) Inhomogeneous eddy current fields. Because of the rapidly rising magnetic field, eddy currents must be carefully watched as, generally speaking, if they occur they cause a phase lag in the field at localized points. Trouble may arise if the flux is forced to cross from one lamination to another, or if the insulation between laminations is locally bad. For a reason which is obscure, difficulty is often encountered in obtaining a sufficiently high resistance between laminations. All metal parts in the magnetic field must of course be either laminated or of sufficiently high surface resistivity. Inhomogeneities due to eddy currents are dealt with in the same way as remanence effects.

(3) Current-carrying conductors near the guiding field. This effect is generally not very serious. Currents flowing into or out of the coils may cause disturbance, and any unbalanced current is dangerous. An effect also to be watched with coils close to the doughnut is termed by Kerst 'flux squirting'. If two coils are run in parallel close together, and if, as is then usually the case, the flux linking one is not the same as the flux linking the other, a constant circulating current, limited at injection only by the low resistance of a coil, will flow through both coils (cf. 'flux forcing', Kerst 1945). This will make the flux squirt towards the doughnut and may cause a horizontal component of magnetic field.

(4) Mechanical unevenness in the assembly of the pole tips. This is clearly a source of magnetic field disturbance.

(5) Finally, saturation of the iron may cause the law of magnetic field to change dangerously, setting an upper limit to the energy reached. This fault does not usually apply to betatrons, but mainly to synchrotrons.

To measure the magnetic field, various devices have been used, the peaking-strip (Bosley, Craggs and McEwan 1947) being invaluable for fields near injection time. To compare fields accurate coils also are used and the voltages induced in them compared one with another (Lasich 1949).

The energies added per revolution are quite small and are of the order of 50 electron volts, so that the electrons traverse the doughnut for the order of a million revolutions.

(*b*) *Vacuum system.* This does not call for anything unusual in most machines. Although the particles circulate several million times in a typical betatron, the gas pressures do not have to be so very low, and 10^{-5} mm. Hg is quite usual. The chamber is made of glass and in one piece, or, if large, porcelain sections joined together with rubber tape are used (Westendorp and Charlton 1945). No special difficulty seems to be encountered with the vacuum system.

However, the inside must be completely lined with an electrical conductor so as to prevent charges from accumulating. It has been reported from several sources that as little as a few square millimetres of uncovered surface will collect enough charge to stop the operation of the machine (Kerst, private communication, Oxford group, private communication).

(*c*) *Injection gun.* Most betatrons (and electron sychrotrons) employ a simple type of electron gun for injection purposes. A flat vertical beam of electrons is produced and is injected usually from the outside at energies ranging from 10–70 kv. The injection energy is often increased with time, and the whole electron beam pulsed. In this way all particles entering the betatron can be made to follow the same path, and the injection oscillations are reduced to a minimum. It is then clearly most important that the timing of this pulse be correct relative to the magnetic field. This is shown by Figure 27, and timing to a few tenths of a microsecond is usually necessary. By this means the output is increased

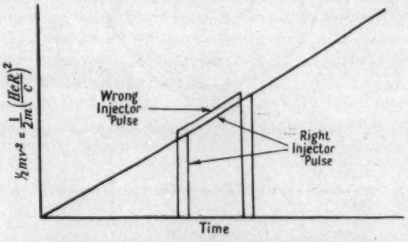


Figure 27. Timing and shape of injector pulse.

several-fold. The use of the contractor coil is also quite common; the reasons for using this have already been summarized. For maximum yields the injector must be placed just near the outside edge of the region of stable magnetic field (i.e. the region where $n=1$). Large currents of electrons are usually used, of the order of 1 amp.

(*d*) *Extraction and X-ray production.* The partial extraction of the accelerated electrons from betatrons has been achieved by two different methods.

By the first method (Skaggs *et al.* 1946) the beam was disturbed by reducing the field around the orbit by means of the contractor coil. This caused the beam to expand and to move into the unstable region of magnetic field, where it entered a region of much reduced magnetic field formed by laminated iron shields called the 'peeler'. By this means a usable number of electrons was extracted. The other method is illustrated in Figure 28 (Gund and Reich 1949). The particle paths are allowed to expand in radius until they pass through a thin aluminium wire. This scatters them without appreciable energy loss into a cone of half angle 1·15°. This causes most of them eventually to reach an electrostatic deflecting system, similar to that of a cyclotron, by the paths shown in the diagram. It is claimed that 75% of the beam is extracted in this way, the electrons eventually emerging through a 0·02 mm. copper window at 4·6 MEV. energy. The fundamental difficulty in extraction of particles from such a machine is due to the fact that the particle period of revolution is so short. It is usually of the order of

0·01 μsec., and this makes it virtually impossible to use any system of pulsed deflection which would be sure to bring the particles out in about one revolution.

The usual method of using the electron beam is to generate x-rays by making the electrons strike a target placed on the inside edge of the doughnut or on the electron gun. The target is in the form of thin tungsten or platinum. The electrons are made to spiral in and hit the target by saturating the centre core at a certain field strength, thus reducing the accelerating field at the orbit. A narrow

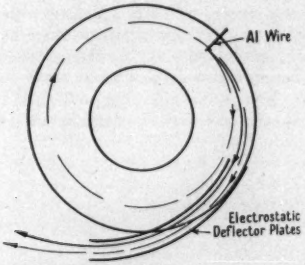


Figure 28. Beam extraction from betatron (Gund and Reich 1949).

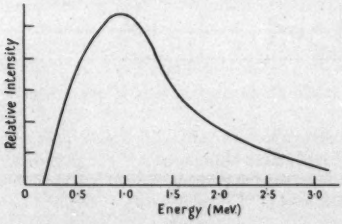


Figure 29. X-ray spectrum from betatron (Lasich and Riddiford 1947).

beam (cone angle $\sim m_0c^2/E$) of x-rays is produced which emerges in the direction of the electrons before they hit the target. Several workers have analysed the spectrum thus obtained, one result being shown in Figure 29 (Lasich and Riddiford 1947). It is different from the expected spectrum for thin targets; this is accounted for by assuming that there is a greater energy spread in the beam than theory predicts. Other investigators find theory satisfactorily confirmed (Koch and Carter 1950).

(iv) *Operation*

Most betatrons operate at the local mains frequency, which means that a pulse is produced every mains cycle. The outputs obtained vary considerably, a figure of 10*R* per min. at 1 m. being an average value. Bess and Hanson (1948) have determined the number of particles accelerated in Kerst's 20 MeV. betatron, and find it to be of the order of 10^{12} particles. To obtain the most from the machine it is often possible to vary the phase of the accelerating flux relative to

the guiding field as this alters the position of the stable orbit. A peculiarity noted by several investigators is the occurrence, under certain conditions, of two output pulses of X-rays separated by the order of a microsecond (Oxford group, private communication, Elder *et al.* 1947).

(v) *Future Possibilities*

The betatron as the most economical means of accelerating electrons to high energies has for many purposes been replaced by the synchrotron; however, its importance in solving the major problems associated with the electron synchrotron cannot be overstressed. The largest betatron actually working is the one at the General Electric Laboratory, Schenectady, and this produces a peak energy of 100 MEV. (Westendorp and Charlton 1945). Kerst is building a large machine, and design figures show that an energy of 300 MEV. should be available.* This is likely to be the largest betatron to be made. Special circuits are required to increase the acceleration rate of the electrons as they reach high energies ($\sim$100 MEV.), since they then begin quite seriously to lose energy by radiation. The rate of loss of radiation is proportional to $(E/E_R)^4/R$, where E is the total energy, E_R the rest energy and R is the radius of the orbit (Iwanenko and Pomeranchuk 1944, Blewett 1946).

It should be mentioned that the induction accelerator is of no economic use in accelerating protons, as for the same energy as electrons one would require a magnet of 40 times the radius.

§ 5. ELECTRON SYNCHROTRON

(i) *General Principles*

The success of the betatron as an electron accelerator encouraged thought on a machine which would not require the massive magnet to supply the changing flux which simply accelerated the electrons. Instead, acceleration was to be achieved by an R.F. electric field as in a cyclotron. Figure 30 illustrates the idea. A magnet which guides the particles inside a doughnut is still necessary. A single acceleration gap is used as an example. The radio frequency must be locked to the average electron frequency, so that on every revolution the electrons experience an acceleration. Provided electrons are moving with energies of 2 MEV. or more they have velocities which are within a few per cent of that of light. Hence beyond this energy the electron velocity and its frequency of revolution remain sensibly constant, and therefore the R.F. acceleration frequency can likewise remain constant. It is in this region that electron synchrotrons operate.

Such a machine was proposed towards the end of the war independently by Oliphant, Veksler and McMillan (Oliphant 1947, Veksler 1944 and McMillan 1945). It was called by McMillan a synchrotron because of a vital property of the particle motions, which resembled the operation of a synchronous motor.

The important question in such a machine is how accurately the accelerating R.F. voltage needs to be related to the rate of rise of magnetic field. As in the betatron, the electrons require on the average a definite amount of energy to be given to them per revolution in order that they may remain inside the doughnut. Whether electrons arriving at the accelerating gap and receiving a different amount

* *Note added in proof.* This machine has now come into operation. A beam at 315 MEV. has been obtained with a radiation output of 1,000 r/min. at 1 metre behind $\frac{1}{4}$" of lead (Kerst *et al.* 1950).

of energy would be lost from the system was dealt with by Veksler and McMillan; they showed independently that the electron motions are stable with respect to the phase at which they arrive at the acceleration gap. The theory is similar to that applied to the synchro-cyclotron. There is a certain phase, called the 'stable phase', ϕ_s (see Figure 31), such that electrons arriving at the gap at that phase would receive just the right amount of energy to keep them circulating on approximately the same circle (the stable orbit). If the particles arrive earlier,

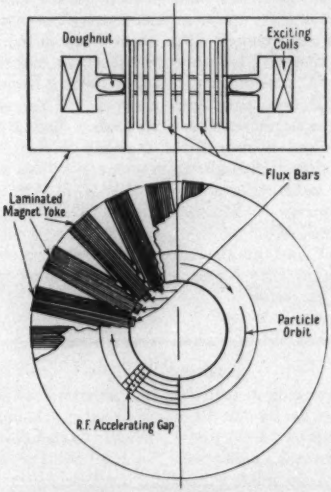


Figure 30. Electron synchrotron.

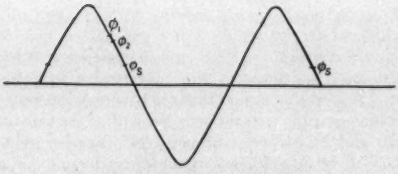


Figure 31. Phase oscillations.

say at ϕ_1, then they receive more than the required amount of energy. This causes them to follow a circle of larger radius and so, because their velocity is constant, arrive later at the gap after the next revolution. This process is repeated and the particle is found to oscillate in phase about the stable phase. A most important property of this oscillation is that it is damped in time at a rate proportional to $E^{-1/4}$, where E is the total energy. These oscillations are called 'phase oscillations'. Thus, provided the R.F. voltage is made larger than the minimum

energy to be added per revolution, there is no strict connection at all between R.F. volts and rate of change of magnetic field. Accompanying the phase oscillations there are oscillations of the particles about the stable orbit in the doughnut. For a general discussion of the properties of these particle motions see Bohm and Foldy (1946), Dennison and Berlin (1946), Frank (1946) and Gooden, Jensen and Symonds (1947).

Some of the phase-oscillation equations may be of some interest (Bohm and Foldy 1946).

Phase equation.

$$\frac{d}{dt}\left(\frac{E_s}{\omega_s^2 K}\dot{\phi}\right)+\frac{r_s L_s' c^2}{2\pi(1-n)K v_s 2\omega_s}\dot{\phi}+\frac{eV\sin\phi}{2\pi}=\frac{eV}{2\pi}\sin\phi_s. \quad \ldots\ldots(10)$$

Energy–radius relation.

$$\frac{\Delta E}{E_s}=(1-n)\frac{v_s^2}{c^2}\frac{\Delta r}{r_s}, \quad \ldots\ldots(11)$$

where E = total energy of electron, E_s = energy of particle on synchronous orbit, ω_s = angular frequency of particle on synchronous orbit, v_s = velocity of particle on synchronous orbit, r_s = radius of synchronous orbit, ϕ = phase of the radio frequency when the electron reaches the accelerating gap, ϕ_s = synchronous phase, V = voltage amplitude of radio frequency, n = magnetic field index defined by $H_z=H(r_0/r)^n$, $K=1+\left(\frac{n}{1-n}\right)\left(\frac{c}{v_s}\right)^2$, $L_s'=\partial L_s/\partial r$, where L_s is the radiation energy loss per turn on the synchronous orbit. Equation (10), neglecting the L_s', is that of a pendulum of moment of inertia $I=E_s/\omega_s^2 K$, restoring torque $(eV/2\pi)\sin\phi$ and acted on by a constant torque $(eV/2\pi)\sin\phi_s$. The rest point of the pendulum is at $\phi=\phi_s$, and beyond $\phi=\pi-\phi_s$ the pendulum executes continuous circular motion. This pendulum analogy was first pointed out by McMillan (1945) and has proved very useful.

The occurrence of the dangerous oscillations arising from coupling between vertical and horizontal oscillations in the synchro-cyclotron when $n=0{\cdot}2$ (see § 3) has led Courant (1949) to make a theoretical investigation into the possibility of a similar danger being present in synchrotrons. He finds that for $n=0{\cdot}75$ a coupling between the second harmonic of the radial oscillation and the particle frequency can give rise to disastrous results. However, this will only occur if the field is very homogeneous circumferentially and if n is very nearly equal to 0·75.

It was mentioned above that electron synchrotrons operate in the region of energies from 2 MeV. upwards (except for one case proposed by Crane (1946)). In order to pre-accelerate the electrons, a betatron method of acceleration is employed (Pollock 1946). Some iron 'flux bars' are made to link the orbit; these collect considerable flux as the field begins to rise, and thus the induced betatron-type field is produced in the doughnut. After the particles have reached an energy of about 2 MeV. the flux bars saturate and the R.F. volts are switched on. In this way operation is switched over to full synchrotron operation. During the change-over the phase oscillations are set up, and with them the accompanying radial oscillations. However, because the R.F. voltage increases with time, it is possible to collect effectively all the electrons already circulating in the betatron phase. The theory of this process was studied by Bohm and Foldy (1946), Kaiser (1949) and Goward (1949).

The synchrotron was first experimentally demonstrated by Goward and Barnes (1946), who used a 4 MeV. betatron, ran the guide field up higher, and accelerated the electrons from 4 MeV. to 8 MeV. by the synchrotron principle. This work was soon followed by others, notably Pollock and his collaborators (Elder *et al.* 1947) at the General Electric Laboratory, Schenectady. There are now numerous electron synchrotrons in operation, the two largest being at Berkeley (300 MeV.) and at Cornell (200–300 MeV.) in the United States of America. There are at least four others of similar size in the course of construction and many smaller ones. Notable work in the development of the smaller machines has been performed at the Atomic Energy Research Establishment, Malvern.

(ii) *Physical Construction*

As has been mentioned several times already, most of the major problems encountered in synchrotron operation are the same as those met with in the betatron. The magnetic field especially offers problems at injection time. The injection systems in all electron synchrotrons which are operating employ the betatron principle, and so their problems are the same. Some new specific difficulties arise due to the different physical design of magnet which is now possible, and to a small extent to the increased size of the machines generally.

(*a*) *Magnet.* The most economical type of magnet design is the one illustrated in Figure 30. This is termed a 'cee' type magnet with outside return yoke. The same considerations as in betatron magnet design still apply. Especially important in this case is the transferring of the flux from the yoke segments to the pole tips. In doing this the flux must cross through laminations. This causes considerable local phase shifts in the magnetic field. However, because of the many yoke segments involved it is an easy matter to place corrector coils around them to help smooth out the inhomogeneities. With large electron synchrotrons energy loss in the magnet becomes large, and instead of allowing the magnet and storage condensers to oscillate at the mains frequency it is often the case that the magnet is allowed to oscillate for one cycle, is then switched off and, after a number of cycles, switched on again. With this type of oscillation new difficulties can arise, the general cause being transient effects when the current is first switched on. For example, it takes a certain finite time for the flux to enter the iron uniformly, and eddy currents will take a finite time to reach a steady state. Furthermore, it is more difficult in this type of operation to ensure that the iron is taken through the same magnetic cycle each time. To avoid these difficulties it is possible to put a negative bias on the whole magnet so that the injection field is reached later than without the bias. This is done at Berkeley (private communication), where a negative field of 16 gauss is applied. It may also be necessary to place coils across the pole-tip faces so that the value of n can be adjusted at injection, as transient effects may cause a change in the magnetic field shape. Switching the magnet off at zero or near zero current is also a problem. As in some betatrons, extra coils around the flux bars are provided so that the relation between flux and magnetic field (equation (7)) can be adjusted.

(*b*) *Vacuum system.* The vacuum system in most electron synchrotrons does not usually offer any serious problems. In small synchrotrons an all-glass system is used and in larger machines porcelain or glass segments are joined together by a variety of simple means to make a complete doughnut. The conducting coating is a thin coating of metal. A special segment of the vacuum

system is usually employed for supplying the R.F. accelerating voltage. Operating pressures, as for betatrons, are from 10^{-5} to 10^{-6} mm. Hg, and the system is continuously pumped.

(*c*) *R.F. system.* Since the radio frequency required remains constant it is not difficult to obtain a source of sufficiently high R.F. power. Quite considerable voltages are required for the high-energy machines, as electrons radiate energy at quite a large rate when moving in a circular path at nearly the velocity of light. The problem of radiation has been considered by Iwanenko and Pomeranchuk (1944), McMillan (1945), Schiff (1946) and Blewett (1946). For the 300 MEV. machines the rate of radiation is about 2,000 ev/rev. This energy loss must be made up from the R.F. system, and this increases the voltage required. The greatest difficulty experienced with the R.F. system appears to be in obtaining a satisfactory acceleration gap. A common practice is to use a special high-frequency porcelain and to construct a resonator out of a section of the vacuum box as in Figure 32. If a conventional type magnet (Figure 30) is used, space must obviously be conserved in the vertical direction. Thus a resonator of high Q but small vertical dimensions is required and the losses must not be too high in the porcelain.

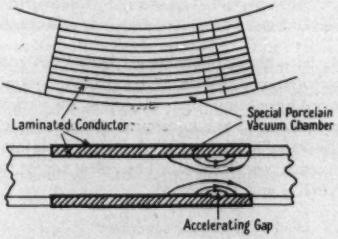


Figure 32. Synchrotron accelerating cavity.

Furthermore, the metal forming the resonator walls must be laminated to reduce eddy currents to negligible proportions. These requirements need very careful design. The powers required are of the order of 1–10 kw. and the R.F. volts vary from a few hundred to 10,000.

(*d*) *Injection.* Injection systems in electron synchrotrons do not differ essentially from those employed in betatrons. Injection energies up to 100 kw. are employed—in one case 1 MEV. (Crane 1946). One extra control which is now necessary but which occurs after injection is complete is the timing of switching on the R.F. system. This timing is found to be critical to within a few microseconds and the radio frequency itself must also be adjusted for maximum output, as also the R.F. voltage.

(*e*) *Extraction.* So far no beam from a synchrotron has been extracted although several workers are carrying out experiments to that end. Schemes similar to those discussed for the betatron will be tried, as well as other methods (Goward and Dain 1947, Clark, Getting and Thomas 1946, Crittenden and Parkins 1946). In the cases of operating synchrotrons the electrons are made to impinge upon an internal target as in most betatrons, and experiments are performed on the beam of X-rays so produced.

(iii) *Operation*

In getting a machine into operation sensitive radiation detectors are found to be invaluable. For example, if a sensitive detector is placed on the inside radius of the doughnut, as soon as the beam gets out of control it will spiral in and record itself on the detector. By reducing the rate of rise of magnetic field (and so overcoming both remanence and eddy current troubles) until a beam is picked up on the detector, the various and many adjustments can be made to the beam.

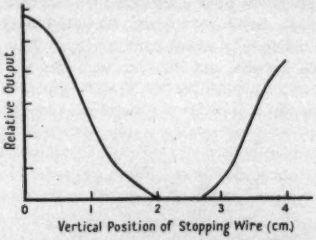


Figure 33. Effect on synchrotron output of vertical aperture.

Outputs comparable with those of betatrons have been recorded (Elder *et al.* 1947). Several investigators (Pollock *et al.*, private communication, Goward, private communication) have found that restricting the vertical aperture reduces the output of the machine until at a minimum aperture there is no output at all. Such behaviour is illustrated in Figure 33. It is clear therefore that most machines operate with small vertical oscillations. This effect might well be explained by gas scattering (see § 6 (ii *c*) and Figure 38).

(iv) *Future Possibilities*

Besides the conventional type of machine which has been discussed above, a number of variations have been suggested, all concerned with the construction of the magnet.

Crane (1946) has proposed a 'race-track' type of device as shown in Figure 35 (*b* and *c*). This has the advantage of having regions free from the magnetic field in which the injection system, R.F. system, extraction system and vacuum pumps can be conveniently placed. It is very suitable for proton synchrotons where betatron injection is not employed. Crane proposes to inject electrons at 1 MEV. and to frequency-modulate the R.F. over a few per cent in order to follow the electron frequency. The theory of the (race-track) has been investigated by Dennison and Berlin (1946), Serber (1946) and Blachman and Courant (1949). There is very little difference in its operation from that of a circular synchrotron.

Another variation is the use of a magnet made of current-carrying conductors only, without the core of iron. This suggestion has come from Kaiser and Greanias (1946), Oliphant (1947) and Blewett (1947). It has the advantage that higher fields can thus be obtained with a consequent possible saving in stored magnetic energy. The strong magnetic forces acting on the conductors make the problem of accurate placing a difficult one. Such a machine, designed to produce 300 MEV. and possibly 1,000 MEV. electrons, is under construction at the General Electric Laboratories, Schenectady.

The upper limit of electron acceleration by cyclic accelerators is met when the rate of energy loss by radiation becomes too high. Already for quite moderate-energy machines (~70–100 MeV. at Schenectady) the radiation is partly in the visible spectrum and quite an intense point of light is visible. It is generally reckoned that about 1,000 MeV. will be the upper limit of electron synchrotron acceleration, when for a machine of 15-foot radius an energy loss of about 20,000 ev. per revolution occurs.

§6. THE PROTON SYNCHROTRON

(i) *General Principles*

The machine which offers the possibility of reaching the highest energy of all is the proton synchrotron, variously called the 'Bevatron' (American West Coast) and the 'Cosmotron' (American East Coast). This was proposed by Oliphant in 1943 (Oliphant 1947) and was inherent in McMillan's suggestion of 1945. Considering that it was only postponing matters to avoid the difficult question of proton acceleration, Oliphant boldly decided in 1946 to build a 1,300 MeV. proton synchrotron. This machine is now entering its final stages of construction at Birmingham University, England. Since 1946 two other machines have been proposed and started, both in the U.S.A.—one at Brookhaven (Livingston 1948, White 1949) and one at Berkeley (Brobeck 1948). The first is for 3,000 MeV. and the second for 6,000 MeV. In order to carry out injection experiments etc. for the large machine at Berkeley, a quarter-scale model was constructed and successfully operated in the summer of 1949 (private communication). This then was the first proton synchrotron to work. It delivered protons of 5 MeV. energy, and produced valuable information for the larger machine. It will hereafter be referred to as the $\frac{1}{4}$-scale model. The four machines known to be under construction are compared in the Table.

Machine.....	Birmingham	Brookhaven	Berkeley $\frac{1}{4}$-scale	1st stage	2nd stage
Max. proton energy (MeV.)	1300	2960	5·9	1650	6440
Orbit radius (cm.)	450	900	350	1450	1524
Acceleration time (sec.)	1	1	0·3	1·5	1·85
Energy added per rev. (ev.)	220	600	43·5	810	1750
Magnetic field at injection (gauss)	227·3	320	326	315	299
Max. mag. field (gauss)	15000	13800	1000	5800	16000
Injection energy (MeV.)	0·5	4	0·625	10	10
Acceleration aperture (cm.)	35×10	67×20	105×30	180×60	120×30
Proton frequency range (Mc/s.)	0·34–9·6	0·39–4·2	0·39–1·2	0·38–2·43	0·36–2·46
Deuteron frequency range (Mc/s.)	0·24–7·8	0·28–3·86	0·28–3·86	0·27– ?	0·26– ?
Frequency tolerance at injection (%)	0·5	?	0·5	0·3	0·2
Peak energy gain per rev. (ev.)	600	?	510	9820	22300
Repetition rate of acceleration (per min.)	6	?	16	10	10
n.....	0·7	0·6	0·6	0·6	0·6

In principle the proton synchrotron is just the same as the electron synchrotron except that now the radio frequency must be varied a considerable amount in

order to follow the increase in speed of the protons as they are accelerated; protons do not begin to approach the speed of light until they have energies of about 1,000 MeV. Another difference in principle is that protons are injected into the synchrotron as a synchrotron and not a betatron. It is not economic any longer to use the betatron principle for the initial acceleration. This means that the injection process cannot be compared closely with that of an electron synchrotron. The motions of the particles in a proton synchrotron were investigated by Gooden, Jensen and Symonds (1947) and Twiss and Frank (1949).

In the non-relativistic region the particles execute phase oscillations as in the electron case, but to a first approximation there is no damping of the oscillations. The factors which affect the damping can be readily controlled in principle, so that anti-damping can be avoided. There are three ways in which the damping can be made finite: (*a*) by increasing the R.F. voltage during acceleration, (*b*) by making the magnetic field rise as t^s where $s>1$, (*c*) by sloping the accelerating electrode gap (Gooden *et al.* 1947). Variations in the radio frequency can also affect the damping rate one way or the other, but only to a small extent. As the particles approach the velocity of light significant finite phase damping occurs.

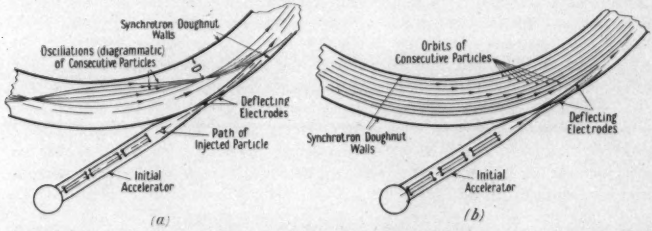


Figure 34. Injection systems for proton synchrotrons.

The motion of the protons when injected into the machine is now totally different from the betatron case. There are numerous schemes for the injection process following those used in the betatron. Figure 34 illustrates the two types most favoured. A high-tension set of some kind accelerates the protons to energies of about 1 MeV. In the first system the energy is kept constant and the particles are guided into the machine by electric and magnetic fields, and directed appropriately. The magnetic field is rising but the R.F. voltage has not yet been switched on. During this time the particles will spiral in as the instantaneous orbit shrinks. At the same time the betatron-type oscillations of particles arriving later in the doughnut will increase. By the time the mean orbit has reached the centre of the chamber there will be particles with amplitudes of oscillation between zero and D (Figure 34). All others have been lost to the walls. At this moment the R.F. volts are switched on. Particles are then further selected. The number selected is determined by their lying between the stable limits of phase and/or by having their resultant radial-oscillation amplitudes lying inside the stable regions of magnetic field (Gooden, Jensen and Symonds 1947, Twiss and Frank 1949). The radial oscillations are made up of the betatron-type oscillations and the synchro-radial oscillations which accompany the phase oscillations. If the R.F.

voltage increases appropriately, particles in a certain region can be accepted over the whole phase range of 2π radians, provided their radial oscillation amplitudes are less than the half chamber width. This is similar to the process of betatron–synchrotron change-over and was demonstrated experimentally and theoretically by Kaiser in an electron synchrotron (Kaiser 1949). It is not of very much help in a proton synchrotron as the main limitation to acceptance of particles is set by the radial oscillations. It must be remembered that as the R.F. voltage comes on slowly, the whole bunch of particles continues to move inwards in the rising magnetic field so that many which are not otherwise lost will hit the walls.

The second injection process is the same as the one outlined, except that the energy of the incoming protons is increased in time. All protons entering the synchrotron have the same radius of curvature in the field, and consequently their betatron-type oscillations are reduced to a minimum. The R.F. volts are switched on as before except that the timing is not now critical. This scheme grew out of the one usually employed in betatrons. It gives a theoretical yield of between three and four times the first scheme. The theory of these schemes and description of the particle motions is given by Gooden (1948, thesis) and Twiss and Frank (1949). Other schemes have been proposed (Oliphant, Gooden and Hide 1947).

Some general formulae are given below. Much of the useful information, especially on injection, can only be obtained by numerical methods, and the reader is referred to the original papers for the necessary formulae (Gooden, Jensen and Symonds 1947 (1), Twiss and Frank 1949 (2)).

Phase Damping for Small Oscillation Amplitudes

(1) *Non-Relativistic Case.* If the magnetic field is varied so that the stable phase ϕ_s is kept constant, or if V is very large, then

$$\phi = \phi_s + GV^{-1/4}\sin(2C^{1/2}t + \alpha), \qquad \ldots\ldots(12)$$

where
$$C = \frac{Vcnp\cos\phi_s}{(x+1)Pr_s^2(1-n)}, \qquad \sin\phi_s = \frac{2\pi x P r_s^2}{Vc}.$$

(2) *Extreme Relativistic Case.*

$$\phi = \phi_s + G_1 t^{(1-2x)/4}.\sin(2C_1^{1/2}t + \alpha_1), \qquad \ldots\ldots(13)$$

where
$$C_1 = \frac{Vpc\cot\phi_s}{P(1-n)r_s^2}.$$

(3) *General Case.*

$$\text{Amplitude of } \phi \propto t^{-g(1+s)/\{4(2+g)\}}. \qquad \ldots\ldots(14)$$

Alternatively,

$$\text{Amplitude of } \phi \sim \left[\eta\left(\frac{\dot{B}_s B_s}{\omega_s}\right)\cot\phi_s\right]^{-1/4}. \qquad \ldots\ldots(15)$$

(4) *Damping due to Frequency Variations.*

$$\text{Amplitude of } \phi \propto t^{(1+x)A}, \qquad \ldots\ldots(16)$$

where
$$A = f(t) - \frac{(1+x)}{t}\int_0^t f(t)\,dt + \frac{4(1-n)t}{n(1+x)}\frac{\partial f(t)}{\partial t}.$$

(5) *Damping due to Electrode Shaping.*

Amplitude of $\phi \propto t^{-D}$ where

$$D = \frac{xp \tan\theta \cot\psi/2}{8(1-n)}. \qquad \ldots\ldots(17)$$

θ is the angle between the radius vector and the electrode face, and ψ is the angle subtended by the 'cee' at the centre of the system.

The nomenclature not already introduced in § 5 (i) is as follows: B_s = magnetic field at synchronous orbit: it increases in time as $B_s = Pt^x$; P is as defined in B_s; p = ratio of R.F. to particle frequency and is a positive integer; x is a positive index determining the way the magnetic field increases with time (see B_s); $f(t)$ is the fractional departure from the frequency law required to keep the particle on a stationary orbit, e.g.

$$\text{R.F.} = \frac{v_s p}{2\pi r_s}(1 + f(t));$$

$g = \dfrac{mc^2 - m_0c^2}{m_0c^2}$; G, G_1, α, α_1 are arbitrary constants; $\eta = \dfrac{(m/m_0)^2}{(m/m_0)^2 - (1-n)}$.

The voltage V must be proportional to t^{n-1} in order to keep ϕ_s constant.

(ii) *Physical Construction*

(*a*) *Magnet.* Besides the differences in principle between electron and proton synchrotrons, quite large differences arise in methods of construction. These differences are due to the increase in size of the components more than to their functional uses. The main reason for building a proton synchrotron rather than a synchro-cyclotron is that the former is very much cheaper than the latter. However, it is still a very expensive piece of equipment, and the most economical means of construction must be adopted. This has caused a radical departure from electron synchrotrons in the matter of magnet construction and excitation.

The cost of a magnet 30 feet or more in diameter made from thin 0·014 in. transformer laminations would be prohibitive quite apart from the expense of the excitation equipment. For this reason all proton synchrotrons under construction have a magnetic field rising to its maximum value of approximately 15,000 gauss in approximately 1 second instead of about 1/200 second as in the case of the electron synchrotron and betatron. This means that the magnet can be made from quite thick laminations of mild steel, usually $\frac{1}{2}$ in. thick. Secondly, it means that banks of capacitors are now no longer the most economical means of storing the energy required by the magnet. Rotating machinery is preferred either in the form of a D.C. generator and flywheel or of an alternator and mercury rectifiers. The magnet is pulsed by closing a switch, the current rises, the voltage across the magnet is reversed and the current is brought back to zero, when the switch is opened (see Figure 36). The cycle is repeated about every 10 seconds.

The types of magnet under construction are shown in Figure 35 (*a*, *b*, *c*). They all have their various advantages and disadvantages. From the point of view of efficiency of stored energy they improve from left to right. The race-track type has the advantages already mentioned. Its disadvantages are small and consist in slightly increased radial oscillation amplitudes and in the greater difficulty of aligning all four segments satisfactorily. The Birmingham and Brookhaven types have an advantage in that the vacuum chamber is easily accessible

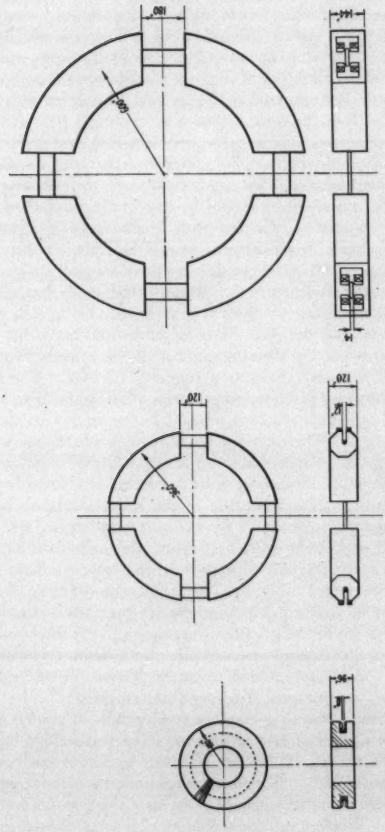


Figure 35. Magnets of the three proton synchrotrons being built.

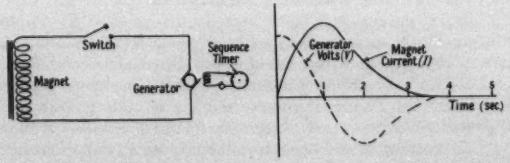


Figure 36. Typical operation of proton synchrotron magnet and D.C. generator.

at all points. A trouble which arises in all such magnets, but which is especially marked in the Brookhaven type, is the following. When the flux density reaches high values—about 10,000–15,000 gauss—the lines of force will lie roughly equally spaced. The permeability of the iron is reduced at these high fields, and thus the M.M.F. around the short path in the iron will be considerably less than that around the outside path in the iron. This may cause the field at the outside of the gap to be reduced appreciably less than on the inside, and make the value of n greater than 1. This could well set the upper limit for the energies obtainable. It is proposed at Brookhaven to place conductors along the pole faces, into which appropriate heavy currents will be forced to correct the field shape at high flux densities. Another possible difficulty with these large-lamination magnets, operated as they are, comes from the presence of the transient effects mentioned in the previous section. All other considerations of magnet design discussed in §5 also apply to proton synchrotrons. Because the radio frequency must be accurately related to the magnetic field, it is necessary to be able to detect any divergence from the relation desired. The methods used vary from one machine to another. As far as is known, Birmingham has the only method for the absolute measurement of the magnetic field as a function of time. For this purpose nuclear magnetic resonance is being employed. The methods used to compare magnetic field and frequency will be discussed in the section on R.F. systems.

(*b*) *Injection System.* The only injection system which has worked so far employs the first type of system discussed above; theory seems to be satisfied (private communications). Injecting into a proton synchrotron is easier in principle than betatron injection because of the independent variable, the R.F. voltage (and if necessary frequency). By not switching on the R.F. voltage until the particles are well inside, the spiralling orbits are made to miss the injection apparatus. However, the relative tolerances on the magnetic field are probably closer because of the larger radius. The oscillation-amplitudes induced by field inhomogeneities or wrong injection-angle are proportional to the radius of the orbit. The size of proton beam, the divergence and the thickness of deflector system will not differ appreciably from an electron system. Hence the relative tolerances on those components which cause oscillations must be closer. The Table shows the injection energies at present contemplated.

(*c*) *Vacuum System.* The long time of acceleration in proton synchrotrons makes important the danger of loss of protons from Rutherford-type scattering with residual gas. A theory of the sort of losses expected has been given by Blachman and Courant (1948). The result of scattering in a focusing field and an infinitely large vacuum chamber appears to be that the protons have a range of induced oscillations. Their amplitude distribution is as shown in Figure 37; the particles effectively concentrate on the outside of a surface, the distance of this surface from the initial orbit being a function of the pressure and varying with time. Thus if the pressure is too high, so that most of the amplitudes of oscillation lie outside the vacuum chamber, most particles will be lost. Blachman and Courant (1948) calculate the loss for the Brookhaven and Birmingham machines for certain valves of the parameters. If the theory of Blachman and Courant is correct, then Figure 38 shows the effect of cutting down the vertical aperture of proton synchrotrons. It is seen that for the dimensions of the Berkeley ¼-scale model the current should begin to disappear for a total aperture of about 4 inches. Since this was found to be the case, gas scattering might well help to explain it.

In any case it is seen that a good vacuum is required in a proton synchrotron. Furthermore, the large radial width of the accelerating space makes a conventional porcelain chamber extraordinarily difficult to manufacture. Many schemes have been proposed for the various machines. A porcelain-type chamber much larger than, but similar to, the electron synchrotron types has been developed for the

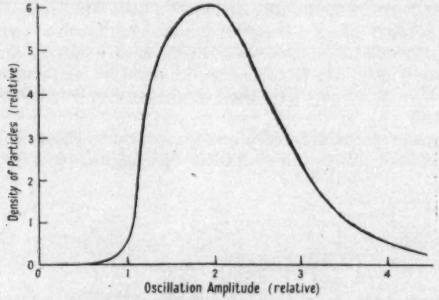


Figure 37. Distribution of particle oscillation amplitudes resulting from gas scattering.

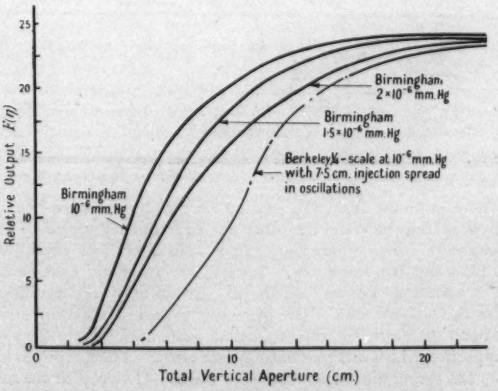


Figure 38. Effect on output of varying vertical aperture (theoretical).
Particle loss by aperture restriction due to gas scattering.
No initial injection oscillations in vertical plane.
$\eta = 0{\cdot}67$. $R = 450$ cm. A = variable. $\epsilon_r = 200$ ev. $E = 500$ kev.

Birmingham machine. The conducting lining is made from laminated 0·018 in. thick Inconel sheet. The porcelain sections are joined by a double rubber gasket. Designs for the other machines are not known, but will probably involve thin stainless steel attached in some way to the magnet pole faces. A construction

of this kind in the Berkeley model gave a vacuum of 10^{-6} mm. Hg with liquid air (private communication). Since the rate of rise of magnetic field in this model was a factor of 4–5 less than, for example, in the Birmingham and Brookhaven machines, relatively thick stainless steel was possible.

(*d*) *R.F. System.* The requirements for the R.F. systems of the three machines under construction are shown in the Table. It is seen that they are quite similar, with the difference that the Birmingham system has a wider frequency coverage, due both to the lower injection energy and to the intention to accelerate deuterons or α-particles as well as protons. The methods of achieving the desired R.F. voltages differ widely. It is clear that in any case considerable electronic ingenuity will be required. The three methods to be employed will be very briefly outlined.

The system proposed at Birmingham is (very much simplified) as shown in Figure 39 (Hibbard 1950). An oscillation following the basic frequency law is

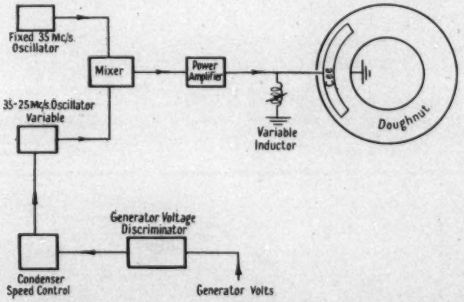


Figure 39. Simplified block schematic of Birmingham proton synchrotron R.F. system.

obtained at low power from the beating of two transmission-line oscillators. One is fixed in frequency and the other has its frequency varied by an accurate rotating capacitor rotating once per second. No direct lock exists between the magnetic field and the frequency. Instead, the magnetic field variation with time will be accurately known (see above). It can only vary significantly if the D.C. generator voltage varies. The generator voltage is therefore automatically monitored and the deviation from a nominal value used to vary the speed of the rotating capacitor, which will be accurately measured. The R.F. power is amplified and finally fed through a transformer into the low-Q circuit of the accelerating electrode. This is an arrangement similar to the accelerating system of the cyclotron. It is simply a drift tube (the 'cee') through which the particles move and for various reasons, including those discussed in connection with cyclotron design in §2(iv), covers about 95° of arc. The need for economy in vertical aperture means that the 'cee' capacitance to earth must be very large, and to increase the impedance level it is roughly tuned mechanically. The variable inductance employed for this purpose consists of a coil plunging into a bath of mercury, which has proved to be very satisfactory.

It is not known whether a final decision has been made on the system to be employed at Brookhaven (Plotkin and Blewett 1949). An induction-type accelerator, suggested by Oliphant, is to be employed, the principle of which is illustrated in Figure 40. This can be used in one of the field-free regions of the magnet. The iron core is constructed of ferroxcube. The basic frequency will probably be generated by the same method as was employed at Berkeley. In this method an inductance with a core of ferroxcube forms part of a stable tuned circuit. A variable direct current biases the inductance, and by effectively varying the permeability of the material alters the inductance. In this way the frequency change is expected to be obtained. By obtaining the D.C. bias from the magnet current, and adding corrections for relativistic effects etc., it is hoped to lock continuously the frequency with the magnetic field. This method was successfully employed in the Berkeley $\frac{1}{4}$-scale model, although the requirements are clearly not nearly so exacting as in a full-scale machine.

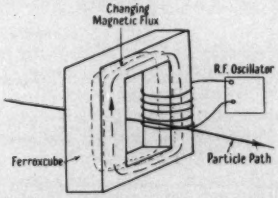


Figure 40. Induction accelerator for proton synchrotron.

The actual accelerator employed in the Berkeley $\frac{1}{4}$-scale model was a 'transit-time' type. This is essentially the same as the Birmingham 'cee' only very much shorter, so that it will fit inside a field-free gap of the race-track.

(*e*) *Extraction.* Various types of extraction system have been proposed. The field-free spaces in the large race-tracks offer a good opportunity of using a pulsed extraction system. Generally speaking, although the energies will be so high, the large radius gives two considerable advantages. The frequency is correspondingly lower, so that a relatively long time is available ($\sim 10^{-7}$ sec.). Also the large radius enables a very long deflector to be used with a correspondingly increased effect. A pulsed electrostatic deflector is to be employed at Birmingham for extraction. Calculations show that by bringing the beam to the outside edge of the stable region of magnetic field 100 kv. on an electrode 12 feet long will extract the beam in a fraction of a revolution.

A stage in the construction of the Birmingham University proton synchrotron is shown in Figure 41.

(iii) *Performance and Future Prospects*

The performance figures of the only proton synchrotron which has so far worked have not been published. In the case of the large machines the outputs will be of a very unusual kind. A large burst of particles will appear for about 10^{-7} sec. only every few seconds. The maximum number of particles which it is estimated can be obtained from the Birmingham machine is between 10^{11} and 10^{12}

per acceleration period. It is clear therefore that special techniques will have to be devised to make the most use of the particles available. The best techniques may well involve changes in the design of the proton synchrotron itself.

The apparent limit at present to the size of the proton synchrotron is the economic one. Modifications of the proton synchrotron are appearing which have economic advantages, for example Oliphant plans to combine an air-core proton synchrotron with a synchro-cyclotron. He calls this device a cyclo-synchrotron (Oliphant 1950).

§ 7. CONCLUSION

In conclusion it may be worth while to consider the factors determining which of the types of accelerators described in this paper are preferable for a given particle energy. The two most important factors are the initial capital cost and the particle output. In Figure 42 are shown the approximate capital costs, related to current British conditions, of the different machines designed to accelerate heavy particles, as a function of particle energy. The absolute values are not to

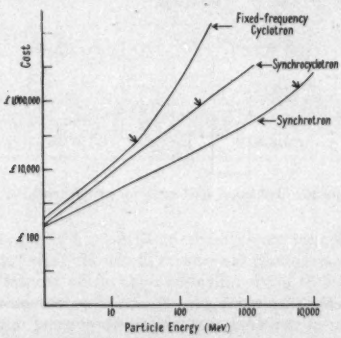


Figure 42. The regions of curves above the short arrows are entirely speculative.

be regarded as particularly accurate, being based on only a few examples, but the differences between different types are great enough for this to be unimportant for our present purpose.

The most expensive item is normally the magnet, and for the purpose of extrapolation we have assumed that the cost of this is proportional to the cube of the linear dimensions, i.e. to the volume of magnetic field for a given field strength. Hence, when relativistic effects are small, so that particle energy E is proportional to the square of the orbit radius, the cost of a magnet of any particular shape is proportional to $E^{3/2}$.

For fixed-frequency cyclotrons the cost of much of the rest of the equipment will also increase more or less proportionally to the volume of the equipment, with the important exception of the radio-frequency power supply. As we saw in § 2, the power requirement varies very rapidly with particle energy, approximately as E^4 for a given machine. In machines of similar shape but of different

scales the ohmic power-losses for a given dee-voltage will decrease as the dee and line dimensions increase, so that for such a range of machines the power input required varies as $E^{7/3}$. We assume that the cost of the R.F. system is proportional to the power, on the ground that increased difficulty of handling higher powers is balanced by savings on various overheads. The cost of the R.F. system is then the main cause for the rapid rise of the curve for fixed-frequency cyclotrons shown in Figure 42. The part of this curve above 25 MeV. for deuterons must even so be taken as a minimum value, since to operate fixed-frequency cyclotrons in this region at all involves the solution of a number of very difficult experimental problems, such as the prevention of high-voltage breakdown. Solutions to these problems may be expensive or even unobtainable.

In the synchro-cyclotron some 30–40% larger energy can be obtained for a given magnet-cost, as the pole-gap can be made much narrower and a larger fall-off of field at the exit-radius can be tolerated. The R.F. system is more complex, and hence more expensive, for small systems, but since there is no critical minimum voltage required there is no reason to suppose that the costs increase with energy faster than those of the magnet.

In the synchrotron the cross-sectional dimensions, and hence the magnetic required energy, do not increase so rapidly as does the radius of the equilibrium orbit. It seems probable therefore that until relativistic increases become serious the cost will, like that of the linear accelerator, rise approximately in proportion to energy. At higher energies the costs will rise rather more rapidly.

The approximate cost as shown in Figure 42 is not by itself of course sufficient to decide the proper machine for a given energy. The current output of a synchrotron is likely to be 100 to 1,000 times smaller than that of a synchro-cyclotron, which in turn gives perhaps 100 times less than a fixed-frequency cyclotron. It is tempting to consider the cost-per-microampere as the proper criterion, but this leads to the striking result that for, say, 300 MeV. deuterons the most economical machine would still be the fixed-frequency cyclotron—at a cost of around £140,000,000. Valuable as 300 MeV. deuterons may be, it may be supposed that the law of diminishing returns should be taken into consideration before embarking on such an instrument. For the purposes of nuclear research the first few particles in a new energy region are of more value than far larger currents ten years later. For exploring new fields there seems no doubt of the immense advantage of the synchrotron. Results obtained by its use will enable a decision to be made as to the value of building a synchro-cyclotron or even a fixed-frequency cyclotron, remembering that the latter machines cost more in time as well as in money.

We may conclude by noting that the real considerations involved may even yet remain undisclosed. It is difficult altogether to avoid the suspicion that the first move made by a modern nuclear physicist is to determine, by means which may only in part be scientific, the maximum amount of money which he can reasonably hope to obtain. Once this primary figure is established, attention can be given to the more detailed matters which, in this review, we have attempted to survey.

ACKNOWLEDGMENTS

The authors are exceedingly grateful to Mrs. J. H. Fremlin and to Mrs. J. S. Gooden for their very extensive help in the preparation of this paper.

REFERENCES

ALLWOOD, H. I. S., 1948, *M.Sc. Thesis*, Birmingham University.
AMALDI, E., and FERRETTI, B., 1946, *Rev. Sci. Instrum.*, **17**, 389.
ATTERLING, H., 1948, *Ark. Mat. Astr. Fys.*, **35**, No. 32.
BARNES, S. W., CLARK, A. F., COLLINS, G. B., OXLEY, C. L., MCCREARY, R. L., PLATT, J. B., and VAN VOORHEES, S. N., 1949, *Phys. Rev.*, **75**, 983 (L).
BESS, L., and HANSON, A. O., 1948, *Rev. Sci. Instrum.*, **19**, 108.
BLACHMAN, N. M., and COURANT, E. D., 1948, *Phys. Rev.*, **74**, 140; 1949, *Rev. Sci. Instrum.*, **20**, 596.
BLEWETT, J. P., 1946, *Phys. Rev.*, **69**, 87; 1947, *J. Appl. Phys.*, **18**, 968, 976.
BOHM, D., and FOLDY, L., 1946, *Phys. Rev.*, **70**, 249; 1947, *Ibid.*, **72**, 649.
BOSLEY, W., CRAGGS, J. D., and MCEWAN, D. H., 1947, *Nature, Lond.*, **159**, 229.
BREIT, G., and TUVE, M. A., 1928, *Carnegie Institute Year Book*, No. 27, 209.
BROBECK, W. M., 1948, *Rev. Sci. Instrum.*, **19**, 545.
BURKIG, J. W., HUBBARD, E. L., MACKENZIE, K. R., 1949, *Rev. Sci. Instrum.*, **20**, 135.
CLARK, J. S., DRAKE, W. W., and TSAO, C. Y., 1949, *Phys. Rev.*, **76**, 162 (A).
CLARK, J. S., GETTING, I. A., and THOMAS, J. E., Jr., 1946, *Phys. Rev.*, **70**, 562 (L).
CORSON, D. R., *et al.*, 1949, *Phys. Rev.*, **76**, 162 (A).
COURANT, E. D., 1949, *J. Appl. Phys.*, **20**, 611.
COWIE, D. B., and KSANDA, C. J., 1945, *Rev. Sci. Instrum.*, **16**, 224.
CRANE, H. R., 1946, *Phys. Rev.*, **69**, 542 (L).
CRITTENDEN, E. C., Jr., and PARKINS, W. E., 1946, *J. Appl. Phys.*, **17**, 444.
DEBRAINE, P., and SIMANE, C., 1948, *J. Phys. Radium*, **9**, 113.
DENNISON, D. M., and BERLIN, T. H., 1946, *Phys. Rev.*, **70**, 58, 764.
DE PACKH, D. C., and BIRNBAUM, M., 1948, *J. Appl. Phys.*, **19**, 795.
ELDER, F. R., GUREWITCH, A. M., LANGMUIR, R. V., and POLLOCK, H. C., 1947, *J. Appl. Phys.*, **18**, 810.
FRANK, N. H., 1946, *Phys. Rev.*, **70**, 177.
FRY, D. W., and WALKINSHAW, W., 1949, *Rep. Prog. Phys.*, **12**, 102 (London : The Physical Society).
GOODEN, J. S., 1948, *Ph.D. Thesis*, Birmingham University.
GOODEN, J. S., JENSEN, H. H., and SYMONDS, J. L., 1947, *Proc. Phys. Soc.*, **59**, 677.
GOWARD, F. K., 1949, *Proc. Phys. Soc.* A, **62**, 617.
GOWARD, F. K., and BARNES, D. E., 1946, *Nature, Lond.*, **158**, 413.
GOWARD, F. K., and DAIN, J., 1947, *Nature, Lond.*, **159**, 636.
GREEN, G. K., SNYDER, H. S., and SMITH, L. W., 1949, *Phys. Rev.*, **75**, 1289 (A).
GUND, K., and REICH, H., 1949, *Z. Phys.*, **126**, 383.
HEYMANN, F. F., 1949, *Phys. Rev.*, **75**, 1951.
HIBBARD, L. U., 1950, *Ph.D. Thesis*, Birmingham University.
IWANENKO, D., and POMERANCHUK, I., 1944, *Phys. Rev.*, **65**, 343.
KAISER, H. F., and GREANIAS, E. C., 1946, *Phys. Rev.*, **69**, 536 (L).
KAISER, T. R., 1949, *Proc. Phys. Soc.* A, **63**, 52.
KAISER, T. R., and TUCK, J. L., 1948, *Nature, Lond.*, **162**, 616.
KERST, D. W., 1941, *Phys. Rev.*, **60**, 47; 1945, *Ibid.*, **68**, 233; 1946, *Nature, Lond.*, **157**, 90; 1948, *Phys. Rev.*, **74**, 503 (L).
KERST, D. W., ADAMS, G. D., KOCH, H. W., and ROBINSON, C. S., 1950, *Phys. Rev.*, **78**, 297.
KERST, D. W., and SERBER, R., 1941, *Phys. Rev.*, **60**, 53.
KOCH, H. W., and CARTER, R. E., 1950, *Phys. Rev.*, **77**, 165.
LASICH, W. B., MUIRHEAD, E. G., and WRIGHT, I. F., 1949, *J. Sci. Instrum.*, **26**, 91.
LASICH, W. B., and RIDDIFORD, L., 1947, *J. Sci. Instrum.*, **24**, 177.
LAWRENCE, E. O., and EDLEFSEN, N. E., 1930, *Science*, **72**, 376.
LAWRENCE, E. O., and LIVINGSTON, M. S., 1931, *Phys. Rev.*, **37**, 1707 (A).
LIVINGSTON, M. S., 1944, *J. Appl. Phys.* **15**, 2, 128; 1946 *Rev. Mod. Phys.* **18**, 293; 1948, *Bull. Amer. Phys. Soc.*, **23**, 2, 28, *Adv. Electronics*, **1**, 269.
LOFGREN, E. J., 1949, *Phys. Rev.*, **76**, 587 (A).
MACKENZIE, K. R., 1948, *Phys. Rev.*, **74**, 104 (L).
MCMILLAN, E. M., 1945, *Phys. Rev.*, **68**, 143.
MCMILLAN, E. M., and SALISBURY, W. W., 1939, *Phys. Rev.*, **56**, 836.

OLIPHANT, M. L., 1950, *Nature, Lond.*, **165**, 466.
OLIPHANT, M. L., GOODEN, J. S., and HIDE, G. S., 1947, *Proc. Phys. Soc.*, **59**, 666.
PICKAVANCE, T. G., ADAMS, J. B., and SNOWDEN, M., 1950, *Nature, Lond.*, **165**, 90.
PLOTKIN, M., and BLEWETT, J. P., 1949, *Phys. Rev.*, **75**, 1288 (A).
POLLOCK, H. C., 1946, *Phys. Rev.*, **69**, 125 (L).
RICHARDSON, J. R., WRIGHT, B. T., LOFGREN, E. J., and PETERS, B., 1948, *Phys. Rev.*, **73**, 424.
ROSE, M. E., 1938, *Phys. Rev.*, **53**, 392.
SCHIFF, L. I., 1946, *Rev. Sci. Instrum.*, **17**, 6.
SCHMIDT, F. H., 1946, *Rev. Sci. Instrum.*, **17**, 301.
SERBER, R., 1946, *Phys. Rev.*, **70**, 434 (L).
SKAGGS, L. S., ALMY, G. M., KERST, D. H., and LANZL, L. H., 1946, *Phys. Rev.*, **70**, 95 (L).
THOMAS, L. H., 1938, *Phys. Rev.*, **54**, 580.
TWISS, R. Q., and FRANK, N. H., 1949, *Rev. Sci. Instrum.*, **20**, 1.
VAN DE GRAAFF, R. J., TRUMP, J. G., and BUECHNER, W. W., 1947, *Rep. Prog. Phys.*, **11**, 1 (London : The Physical Society).
VEKSLER, V., 1944, *C. R. Acad. Sci., U.R.S.S.* (Doklady), **43**, No. 8, 329, and **44**, No. 9, 365.
WALTON, E. T. S., 1929, *Proc. Camb. Phil. Soc* , **25**, 469.
WANG, P. K. S., and WIENER, M., 1949, *Phys. Rev.*, **76**, 1724 (L).
WESTENDORP, W. F., 1945, *J. Appl. Phys.*, **16**, 657.
WESTENDORP, W. F., and CHARLTON, E. E., 1945, *J. Appl. Phys.*, **16**, 581.
WESTENDORP, W. F., and ELDER, F. R., 1949, *Phys. Rev.*, **76**, 445 (L).
WHITE, M. G., 1949, *Phys. Rev.*, **75**, 1288 (A).
WIDERÖE, R., 1928, *Arch. Elektrotech.*, **21**, 387.
WILSON, R. R., 1938, *Phys. Rev.*, **53**, 408; 1940, *J. Appl. Phys.*, **11**, 781.

MESONS

By C. F. POWELL

H. H. Wills Physical Laboratory, The University, Bristol

CONTENTS

PAGE

§ 1. Introduction ... 351
§ 2. Historical survey ... 352
(i) The ' heavy quanta ' of Yukawa ... 352
(ii) Discovery of the μ-mesons ... 353
(iii) The fate of the μ-mesons ... 353
(iv) Discovery of the π-mesons ... 356
(v) Nomenclature ... 358
§ 3. Mass and mode of decay of the π-mesons ... 359
(i) The decay of π^+-particles ... 359
(ii) Mass of the π-particles : Method A; Residual range and grain-density ... 361
(iii) Method B; Scattering and residual range ... 364
(iv) Method C; Momentum and residual range ... 365
(v) Method D; Grain-density and scattering ... 368
(vi) Method E; Ratio of the ranges and of the momenta of protons and π-particles ... 371
(vii) The nature of the neutral particle emitted during μ-decay ... 372
§ 4. Properties of the π-mesons ... 373
(i) Charge of the π-particles ... 373
(ii) Life-time of the π-particles ... 373
(iii) Disintegrations produced by π^--particles ... 375
(iv) Capture of π^--particles by particular elements ... 377
(v) Spin of the π-particles ... 378
(vi) β-decay of the π-particles ... 379
§ 5. Properties of the μ-mesons ... 379
(i) Mass of the μ-particles ... 379
(ii) The decay of the μ-meson ... 380
(iii) Nature of the decay particles ... 382
(iv) Life-time of the μ-particles ... 383
(v) Nuclear interactions of μ^--particles ... 384
§ 6. Artificial production of π-mesons ... 387
(i) Excitation curve for the production of mesons ... 387
(ii) Energy spectrum of the mesons ... 389
(iii) Ratio of the numbers of π^+ and π^--particles ... 390
(iv) Experiments with γ-radiation ... 391
(v) Production of mesons by energetic neutrons ... 391
§ 7. Creation of π-mesons by cosmic radiation ... 392
(i) Ejection of π-particles of relatively low energy during nuclear explosions ... 392
(ii) Production of ' showers ' of penetrating particles ... 393
(iii) Nature of the shower particles ... 395
(iv) Distribution in energy of the mesons of the penetrating showers ... 397
(v) Nuclear interactions of the shower particles ... 398
(vi) Multiplicity of production of mesons ... 400
(vii) Production of ' showers ' at great depths underground ... 401
§ 8. Neutral mesons ... 402
(i) Transmutation of neutral particles into quanta ... 402
(ii) Experimental evidence for the existence of neutral mesons ... 404

(iii) Evidence for the production of neutral mesons by cosmic radiation 406
(iv) Mass and energy of neutral mesons in penetrating showers 407
(v) Life-time and frequency of production of neutral mesons 409
(vi) Spin and parity of the π^0-particles 411
§ 9. Evidence for the existence of mesons of other types 411
(i) Evidence for the existence of τ-mesons 411
(ii) Varitrons 414
(iii) Summary of evidence 416
§ 10. Production of mesons in the atmosphere 416
(i) The primary cosmic radiation 417
(ii) Disintegrations of heavy primary particles 418
(iii) Origin of the 'hard' and 'soft' components 418
§ 11. Outstanding problems and perspectives 420
Acknowledgments 422
References 422

§ 1. INTRODUCTION

THE previous Reports in this series which bear on the subject of mesons were written by Heitler (1939) and Peierls (1940), and contained an account of the main features of our knowledge of these particles up to the end of 1939. Since that date, and especially in the past four years, there has been a rapid development of the subject; and to-day, although many points of detail remain to be elucidated, it seems reasonable to suppose that the most important facts—at least about the most common of the different types of mesons—have been established, and that the relationship of the particles to the cosmic radiation is well understood.

In giving an account of the discoveries of the last ten years I propose to begin with a brief historical survey of the main line of development, then to discuss in more detail the properties of the π- and μ-mesons, together with the evidence for the existence of other types, and finally, to give an account of our present picture of the principal processes associated with the creation of mesons which take place as a result of the passage of cosmic radiation through the atmosphere.

A part of decisive importance in the extension of our knowledge of mesons has been played by the development of new technical resources: the method of recording the tracks of charged particles as a result of their passage through photographic emulsions. Apart from the well-known features of the method—its simplicity and flexibility, the 'integrating' property of the emulsion due to its continuous sensitivity, and the direct and detailed insight into nuclear processes which it allows—it has the advantage of providing a remarkably extended time-scale for the study of certain types of transient phenomena. In particular, particles of very short life-time, such as the π-mesons, which commonly decay 'in flight' when moving in the atmosphere, are arrested in a solid material in a time interval more than a thousand times shorter than when moving in a gas. It is thus possible to study the spontaneous decay of the particles when stopped in an emulsion, or their interaction with nuclei, and thus to observe phenomena which it has proved very difficult, or impossible, to record by other methods.

Because of the important rôle which it has played, an account of the results obtained with the new method necessarily forms an important part of the present report. It is an interesting feature of the historical development of the subject that the π-mesons were discovered very shortly after the development of those technical resources necessary for their observation.

§ 2. HISTORICAL SURVEY

2 (i). *The 'Heavy Quanta' of Yukawa*

The idea that there should exist in nature particles with a mass intermediate between that of the proton and that of the electron first appeared, in 1935, as a result of theoretical speculations by Yukawa (1935). It had been known for some years that the forces between the nucleons of a nucleus—the protons and the neutrons of which it is composed—cannot be electromagnetic in origin. The observed degree of stability of the nuclei can only be accounted for if it is assumed that special forces of a new type—the nuclear forces—come into play between the nucleons, forces which we do not meet in other fields.

Further, it had been shown, as a result of accurate determinations of the masses of the nuclei, that the energy which has to be supplied in order to remove a nucleon from a nucleus—the 'binding-energy' per nucleon—does not change widely in passing from the light to the heavy elements of the Periodic Table. It follows that the nuclear forces are of 'short range', that they vary with the distance between two nucleons more rapidly than in the case of forces governed by an inverse-square law.

Proceeding from these facts, and on the basis of a formal analogy described in detail in the previous Reports, Yukawa concluded that free 'quanta' should exist corresponding to the nuclear field analogous to the photons of the electromagnetic field, that they should have a finite rest-mass of about $150\,m_e$, where m_e is the mass of the electron, and that they should be unstable and decay with the emission of an electron, the life-time of the particles being about 10^{-7} sec. Just as the electromagnetic field of an electron can be regarded, in the quantum theory of radiation (see Flint 1939), as equivalent to the emission of photons, and the force between two electrons as due to a mutual emission and absorption of photons, so the forces of nuclear cohesion between nucleons was attributed by Yukawa to the 'virtual' exchange of charged quanta between the neutrons and protons in a nucleus.

The hypothetical 'heavy quanta' of the nuclear field were thus visualized as differing in important respects from photons. In the first place, the finite rest-mass of the particles was a necessary consequence of the short-range character of the nuclear forces. Secondly, since the continuous exchange of 'quanta' between the neutrons and protons in a nucleus was assumed to be accompanied by an exchange of charge, the neutrons becoming protons and vice versa, it was assumed that, at least in some cases, the 'quanta' carry a positive or negative electric charge. And thirdly, the assumption of the instability of the particles was introduced in order to account for the β-activity of radioactive substances, the process of nuclear decay being attributed to the spontaneous disintegration of the charged 'quanta' into electrons and neutrinos.

In addition to the charged quanta, it appeared necessary to postulate the existence of neutral particles of a similar type in order to account for the forces between like nucleons: between neutrons and neutrons, protons and protons (Fröhlich *et al.* 1938). These forces are known to be of the same order of magnitude as those between protons and neutrons; and, if they have a similar origin, the particles involved in the exchange must be electrically neutral. In what follows it will be convenient to refer to these hypothetical 'quanta', both charged and neutral, as Yukawa particles.

The actual emission of the particles from stable nuclei was not to be expected because such processes are energetically impossible, but in the nuclear collisions involving particles of great energy which occur in the cosmic radiation it was anticipated that such a limitation would not be present.

2 (ii). *Discovery of the μ-Mesons*

What appeared to be a remarkable verification of the correctness of Yukawa's speculations was provided by observations of Anderson and Neddermeyer and of other workers, including Street and Stevenson, in experiments on the nature of particles in the cosmic radiation in the years 1936–8. This radiation can be divided, phenomenologically, into two components with different absorption coefficients in matter. The 'soft' component, which is rapidly absorbed in lead, is responsible for the well-known phenomena of the 'cascade showers' commonly studied with Wilson expansion-chambers and counters (Blackett and Occhialini 1933), and was known to be due to photons and electrons. Anderson (1937, 38) showed that among the particles of the 'hard' or penetrating component there were present particles of mass about 200 m_e, which were able, unlike the electrons, to penetrate many centimetres of lead without losing a large fraction of their energy by the well-known process of 'bremsstrahlung'. These particular particles are now commonly referred to as μ-mesons (Lattes *et al.* 1947b), and the latest measurements of their mass give the value of $(215 \pm 5)m_e$.

The view that the μ-mesons were to be identified with the particles of Yukawa was strengthened by observations which suggested that they are unstable and that they decay with a life-time of about $1{\cdot}5 \times 10^{-6}$ sec., a value only about ten times greater than that anticipated by Yukawa. Further, it was found possible to modify some of the original features of the theory, and thus to account for the magnetic moments of the proton and the neutron and for certain features of the states of energy of the light nuclei, notably the ground and first excited states of the deuteron.

These remarkable successes encouraged the hope that the fundamental features of Yukawa's views were substantially correct, that the μ-mesons were to be identified with the Yukawa particles, and that it would be possible to elaborate the theory to provide a firm basis for the description of the main features of the atomic nuclei and of nuclear collision processes. These hopes were reflected in the previous Reports, but already certain serious theoretical difficulties were beginning to emerge.

2 (iii). *The Fate of the μ-Mesons*

The original estimates of the life-time of the μ-mesons were based on the observation that the reduction in the intensity of a stream of these particles in the atmosphere depends not only on the mass (gm/cm^2) of material traversed, but on the length of the path in which the matter is distributed. The penetration of a given mass of air produces a greater diminution of intensity the lower the density. This effect could only be interpreted by assuming that the removal of mesons by interaction with atoms is accompanied by a spontaneous decay of the particles, the latter process becoming increasingly probable the greater the time of flight (see Follett *et al.* 1936, Auger *et al.* 1937).

Decisive support for the view that the μ-mesons do indeed suffer spontaneous decay with the emission of a charged particle of small rest-mass, tentatively

assumed to be an electron, was provided by experiments of Williams and Roberts (1940). These workers obtained two photographs, by means of a Wilson chamber operated in a magnetic field, of penetrating particles which, after passing through a lead plate, reached the end of their range in the gas of the chamber. From the end of the resultant track a fast particle originated which had a specific ionization near the minimum value for a particle with the elementary electronic charge, and of which the momentum, as determined by the observed curvature of the trajectory in the magnetic field, was equal to 70 ± 35 MEV/c.

The observed specific ionization of the decay particle showed that it was ejected with a velocity closely approaching that of light, c. It therefore followed that its kinetic energy E was approximately equal to pc, where p is the momentum of the particle, and that the energy of the particle was of the order of 50 MEV. This result suggested that in the spontaneous decay of the μ-meson the momentum of the emitted charged particle is balanced by that of a single neutral particle of small rest-mass, a neutrino; for in such a case the total energy corresponding to the disappearance of the rest-mass of the parent meson, $\sim$100 MEV., will be shared equally between the two products into which it decays, and the energy of emission of the electron will be unique and equal to approximately 50 MEV. We shall see in a later paragraph that this conclusion was incorrect.

A more direct and precise determination of the life-time of the μ-mesons was carried out by Rasetti (1941), and later, with greater accuracy, by Rossi and Nereson (1942) and Chaminade *et al.* (1944), by the method of 'delayed coincidences'. In principle, the method consisted in detecting, by suitable arrays of counters, the instants of arrival of μ-mesons of the cosmic radiation (Figure 1). The apparatus selects those particles which, having penetrated a thick layer of lead, are brought to rest in a second block of material; and the time interval is measured between the instant of arrest of a μ-meson and the subsequent discharge of another set of counters by the action of the charged particle which is emitted in the process of decay. The distribution in the observed values of the time delay was found to be similar to that expected from the exponential law of decay characteristic of radioactive substances. The mean life-time of the particles, according to the latest determinations by this method, is $2{\cdot}15 \times 10^{-6}$ sec.

Hitherto, the results of most of the new experiments had given support for Yukawa's speculations. It was, however, difficult to understand why experiments with Wilson chambers indicated that μ-mesons interact very rarely with nuclei in passing through matter. Further difficulties began to appear when observations to distinguish the properties of positive and negative μ-mesons were undertaken.

It had been suggested by Tomonaga and Araki (1940) that when a positive μ-meson is brought to rest in a material we must assume that it will be prevented from approaching a nucleus as a result of the Coulomb repulsion between like charges, and will remain free until the instant of its decay. On the other hand, the electrostatic forces will tend to make the negative particles approach nuclei. It was anticipated that if the spin of the particles is equal to that of an electron, they will fall to states of lower energy round the nucleus, the transitions being accompanied by the emission of radiation or of Auger electrons. As a result of the much greater mass of the particles as compared with that of the electron, they will, in their state of lowest energy, be two hundred times nearer to the nucleus than an electron in its 'k-orbit'; and when in this

state they will spend an appreciable fraction of their time in the nucleus itself. In accordance with these anticipations, both Rasetti and Rossi found that only about half the mesons arrested in their apparatus gave evidence of decay.

In order to confirm the view that the μ^--particles do indeed interact with nuclei when arrested in solid substances before they have had time to decay, experiments were therefore undertaken by Conversi, Pancini and Piccioni (1947). Delayed coincidence experiments were made with positive and negative μ-particles, the two types being separated from one another by magnetic deflection before being allowed to enter the recording apparatus. When the particles were brought to rest in iron, it was found, in accordance with anticipation, that the positive

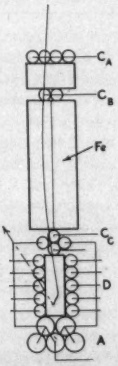


Figure 1. A characteristic modern design of apparatus for the determination of the life-time of +ve or −ve μ-mesons. The two halves of the block of iron are magnetized so that in each case the magnetic intensity is normal to the plane of the paper; but in one half the intensity is directed outwards from the paper, and in the other half into it. This arrangement serves to concentrate particles of one sign. Those particles are selected which discharge a counter in C_A, C_B and C_C, and which do not discharge counters A. Such particles have been arrested in the absorber and the time interval is measured between their instant of arrival and the subsequent discharge of a counter of the set D. The distribution in the values of this time interval gives a measure of the half-value period of the particles—see, for example, Figures 17 and 20. By reversing the magnetic fields, particles of opposite sign can be collected. (*After* Rossi.)

mesons decayed with a life-time equal to that found for the undifferentiated particles, whilst the negative particles disappeared without giving rise to a decay electron.

The above result was at variance with observations by Auger, Maze and Chaminade (1941) which indicated that nearly all the negative mesons arrested in aluminium suffer decay. Although this result was subsequently found to be incorrect (Rossi and Nereson 1942), the experiments described above were extended to the case in which the μ-particles were arrested in a block of material of low atomic number, and for this purpose graphite was chosen. The remarkable result was obtained that in carbon at least a large proportion of the μ^--particles decay with the emission of an electron.

The time occupied by a meson after being reduced to a velocity in the 'thermal' region in falling to the state of lowest energy round a nucleus has been subject

to some discussion and will, it appears, depend on the nature of the material in which it is moving; but it was generally considered to be less than about 10^{-12} sec. (Fermi *et al.* 1947). It was therefore anticipated that such a particle, when brought to rest in solid materials, would approach a nucleus in a time short compared with its life-time. If, further, the μ-mesons are to be identified as the 'heavy quanta' of the nuclear field, such a particle should interact with the nucleus, and the probability of its suffering spontaneous decay should be very small. There was no firm basis for estimating the consequences of this nuclear interaction; whether the disappearance of the particle and the resultant liberation of the energy corresponding to its rest-mass would lead to the emission of γ-radiation, to the disintegration of the nucleus, or to other processes, remained uncertain.

In view of these considerations, the experiments of Conversi *et al.* were of decisive importance, for they appeared irreconcilable with the view that the μ-mesons are to be identified with the Yukawa particles. They indicated that in the process of interaction with carbon, a μ^--particle can reside in its k-orbit for a period of the order of 2×10^{-6} sec. without interacting with the nucleus. The conclusion was drawn that the strength of the interaction between μ-particles and nucleons is less, by many orders of magnitude, than that to be expected if the μ-mesons are identified with the heavy quanta of the nuclear field. Some attempts (Fröhlich 1947) were made to explain the difficulty by assuming that the time occupied by the particle in being brought to rest, and in falling to its state of lowest energy, is much greater than the first estimates had suggested; but whilst some doubt was thrown upon the precise values of this interval, it was made clear that it must be shorter, by some order of magnitude, than the life-times of the particles. A fundamental difficulty therefore remained to be resolved.

Even before the very serious nature of these difficulties had been widely understood, Sakata and Inoue (1946) had suggested that the cosmic-ray evidence appeared to indicate the existence of mesons of two types; and Möller (1941, 1946), on the basis of very general theoretical considerations, had been led to postulate, tentatively, the existence of several kinds of particles of intermediate mass with genetical relationships between them. A little later, Marshak and Bethe (1947), as a result of the contradiction between the large cross-section for the creation of mesons in nucleon–nucleon collisions, and the subsequent very weak interaction of the penetrating particles with nuclei, suggested that the primary products of the nuclear interactions taking place in the atmosphere are heavy mesons, and that these particles decay spontaneously, with a life-time of the order of 10^{-8} sec., to form the mesons commonly observed among the particles of the penetrating component.

2 (iv). *Discovery of the π-Mesons*

It was at this stage in the development of the subject that results were obtained in investigations of the cosmic radiation by means of the more sensitive photographic plates—the 'Nuclear Research' emulsions produced by Ilford Ltd. It was shown by Perkins (1947) and by Occhialini and Powell (1947) that if such plates are exposed at mountain altitudes, the tracks of charged mesons brought to rest in the emulsion can be detected. The masses of the particles were estimated by 'grain-counts' and by studying the deviations in the trajectories

due to multiple Coulomb scattering, and were shown to be of the order of 200–300 m_e. It therefore appeared to be certain that some at least of these particles were identical with the μ-mesons of the penetrating component of the cosmic radiation.

The new plates, whilst greatly superior to the 'half-tone' emulsions which they displaced, were not sufficiently sensitive to record the tracks of electrons moving at relativistic velocities, and it was therefore impossible to observe any fast decay particles emitted by the mesons at the end of their range. On the other hand, it was found that about 10% of the mesons, when stopped in the emulsion, led to a nuclear disintegration with the emission of slow protons, α-particles etc. (see Plates II and III). At the time that this observation was made, it was assumed that this process corresponded to the capture of a μ^--particle by a silver or bromine nucleus in the emulsion: that it was of the same type as that which, in heavier elements, leads to the disappearance of negative μ-mesons before they have had time to decay with the emission of an electron. It was shown later, however, that the mesons which produce nuclear disintegrations are not μ-particles.

Shortly after the observation of the nuclear disintegrations produced by charged mesons, it was discovered in Bristol (Lattes *et al.* 1947a) that about 10% of the mesons, when brought to rest, lead to the emission of a second meson. Further, it was established that the range of the secondary particle is always constant within narrow limits. It was therefore reasonable to assume that it is always ejected with constant velocity, the departures of the individual values of the range from the mean being due to 'straggling'. This strongly suggested that the process corresponds to the spontaneous decay of the parent particle: that two types of mesons exist of different mass, the kinetic energy of the secondary mesons being provided by the disappearance of part of the rest-mass of the heavier, primary mesons. This was subsequently shown to be the case, and it was proved that the secondary mesons are identical with the μ-mesons of the penetrating component of the cosmic radiation. The heavier primary particles were therefore referred to as π-mesons, and their spontaneous transformation, by analogy with the β-decay, was named the μ-decay.

In addition to the isolated tracks of mesons, plates exposed to cosmic radiation were also found to record nuclear disintegrations from which mesons of low kinetic energy were emitted. A large proportion of these 'ejected' mesons, at the end of their range, produce nuclear disintegrations. It was therefore suggested that they are the negative counterparts of the π-mesons, the latter being positively charged and thus unable to interact with nuclei when reduced to low velocities. It was suggested further that the positive and negative π-particles are the primary products of nuclear interactions of great energy occurring in the atmosphere; that being short-lived, however, they decay 'in flight' and thus produce the positive and negative μ-mesons of the penetrating component of the cosmic radiation (Lattes *et al.* 1947b). Further experiments have shown that this view of the origin of the particles, and of their relationship to the cosmic radiation—closely similar to that put forward at the same time by Marshak and Bethe on the basis of other evidence—is substantially correct.

In the three years which have followed the discovery of the μ-meson, great progress has been made in the detailed study of the properties of the particles:

this development forms the subject of the second section of the report. The observations have been facilitated by the discovery, in Berkeley (Gardner and Lattes 1948), of the artificial generation of the π-particles in the bombardment of matter by fast α-particles and protons, and the possibility of making experiments in controlled conditions in the laboratory.

Recent evidence appears to confirm earlier reports from a number of laboratories of the existence of particles of mass about 1,000 m_e which it is convenient to refer to as τ-mesons. The question of the existence of these particles may be of decisive importance for the development of a satisfactory theory of the mesons and of the nuclear forces, and to-day the elucidation of their nature and properties is one of the central problems of nuclear physics.

2 (v). *Nomenclature*

Following the first observations of the tracks of mesons in photographic emulsions, and because of the difficulty of determining the mass and the sign of the charge of the particles, they were classified, phenomenologically, according to the secondary processes observed at the end of their range in the emulsion. A π-meson was defined as one which, when at rest, emits a second meson, μ, with a range of about 600 microns. Later experiments have now confirmed the original view (Lattes *et al.* 1947 b) that these particles are positively charged, and they will be referred to as π^+- and μ^+-particles, respectively.

Secondly, those mesons which produced an observable nuclear disintegration when stopped in the emulsion were referred to as σ-mesons. It is now known that they are exclusively, or almost exclusively, π^--particles which are captured by nuclei. The resulting nuclear disintegrations which produce tracks of charged particles radiating from a centre are commonly referred to as ' stars '. The individual tracks making up a ' star ' are sometimes called the ' prongs '.

Most of the mesons produced no secondary particles which could be observed in the most sensitive emulsions available in the early experiments, and were classed as ρ-mesons. It is now known that in ' electron-sensitive ' emulsions (Berriman 1948)—which record the tracks of particles with the elementary electronic charge even when they are moving at relativistic velocities—about 65% of the ρ-mesons give evidence of decay with the emission of a fast electron. This fraction of the ρ-mesons is due mainly to μ^+-particles formed by the decay of π^+-particles outside the emulsion, and to μ^--particles—formed by the decay in flight of π^--particles—which have been captured by the light elements in the gelatine of the emulsion, carbon, oxygen and nitrogen. In addition, however, a small proportion of these ρ-mesons may be due to π^+-particles which have undergone direct β-decay.

Of the 35% of the ρ-mesons without associated electron tracks, the majority are due to μ^--particles which have been captured by the silver and bromine atoms of the emulsion. The particles interact with these nuclei before they have had time to decay, but no charged particles giving visible tracks are produced in the resulting transmutations. In addition, a small fraction of the ρ-mesons is due to π^--particles which have been captured by nuclei without leading to the production of a recognizable disintegration. Such events will occur if neutral particles alone are emitted in the nuclear transmutation resulting from the capture of the meson, as is the case, for example, when the π^--particle interacts with hydrogen.

In describing the properties of the different types of particles, the terms π^+- and π^-- particles or mesons, μ^+- and μ^--particles or mesons will be employed whenever possible, ρ-mesons or σ-mesons being referred to only when the precise nature of the particles under discussion is ambiguous.

Reference has already been made to the distinction between the 'hard' and 'soft' components of the cosmic radiation, and to the fact that the experiments of Blackett and Occhialini with the counter-controlled Wilson chamber showed that many particles of the 'soft' component often arrive together in the form of the well-known 'cascade showers', which are produced by the multiplication of an original energetic electron or γ-ray. On the other hand, the experiments of Wataghin, Jánossy and others have shown that the penetrating particles of the 'hard' component, most of which we now know to be μ-mesons, can be produced in groups—the so-called 'penetrating showers'. It will be important, in certain sections of this report, to remember the distinction between the 'showers' of the two types.

§ 3. MASS AND MODE OF DECAY OF THE π-MESONS

3 (i). *The Decay of π^+-Particles*

When a positive π-meson is arrested in a solid material it decays spontaneously with the emission of a meson of smaller mass, a μ^+-particle, see Plate I. The distribution in range of 90 μ-particles formed by the decay of π^+-particles stopped in Ilford C2 photographic emulsions is shown in Figure 2, the mean

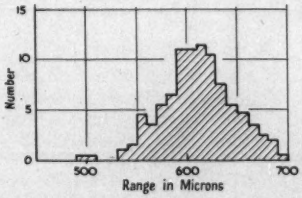


Figure 2. Distribution in the values of the range in Ilford C2 emulsion of 90 μ^+-particles produced by the decay of π^+-particles arrested in the emulsion. The observed spread is consistent with that expected from 'straggling' if the particles are emitted with constant velocity, and proves that a π^+-particle must commonly be moving with very low velocity at the instant of its decay. The mean value of the range is 612 μ, and the corresponding value of the kinetic energy $E\mu = 4{\cdot}15$ MeV.

value being 612 microns. As mentioned above, the observed distribution is consistent with the assumption that the velocity of ejection of the μ-meson is constant within narrow limits. It follows that during the emission of the μ-meson the momentum balance is provided by the ejection of a single neutral particle. The question of the nature of this particle immediately arises: whether it is a photon, a neutrino, or a neutral particle of considerable rest-mass, i.e. a 'neutretto'.

The observed value of the mean range of the μ-mesons in the emulsion allows the momentum and energy of the particle to be determined if its mass is known. When a charged particle passes through matter, it loses energy as a result of interactions of various types with atoms lying near its line of motion: (*a*) by bremsstrahlung, the creation of photons in inelastic collisions with electrons;

(*b*) by inelastic collisions with nuclei in which the latter are excited or disintegrated; (*c*) by ionization loss, due to elastic collisions with electrons; and (*d*) by the production of pairs of electrons in the field of nuclei. With particles of relatively low velocity, the loss of energy by ionization is the dominant process and the others may be neglected.

The rate of loss of energy by ionization per unit length of the trajectory depends on the charge of the particle, Ze, and its velocity, v, but is independent of the mass. For particles with the elementary electronic charge we can therefore write

$$\frac{dE}{dR} = f(v). \qquad \ldots\ldots(1)$$

It follows that the ranges $R_{M,v}$ and $R_{m,v}$ of two particles, each of velocity v, but of masses M and m respectively, are proportional to their masses, viz.

$$\frac{R_{M,v}}{R_{m,v}} = \frac{M}{m}. \qquad \ldots\ldots(2)$$

Suppose that the relation between the mean range of protons for different values of the initial energy has been determined. As is made manifest by the well-known

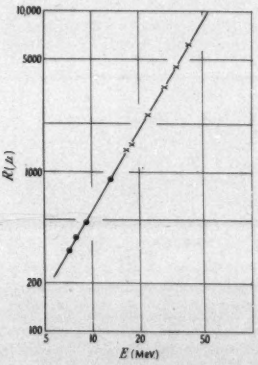


Figure 3. Relation between log R (range in microns) and log E (energy in MeV.) for protons in Ilford C2 emulsion.
● Lattes, Fowler and Cuer (1947).
× Barkas, Bishop and Bradner (1950).
The straight line represents the equation $E(\text{MeV.}) = 0{\cdot}251\ R(\mu)^{0{\cdot}581}$.

Bragg curve for α-particles, the specific ionization of a particle increases rapidly as its velocity decreases, and the relation between the range and energy is therefore not linear.

It has been found that for protons, the relationship can be expressed with a high decree of accuracy by the equation

$$E = kR^n, \qquad \ldots\ldots(3)$$

k and n being constants which must be determined for the particular type of emulsion employed (see Figure 3). From equation (2) it follows that for a particle with a mass $\mathcal{M}$ times that of the proton the relation between range and energy can be written

$$E = k.\mathcal{M}^{1-n}.R^n. \qquad \ldots\ldots(4)$$

The range–energy relation for particles of any assumed mass can therefore be deduced from the corresponding curve for protons, and the energy and momentum of the μ-meson can be determined for any assumed value of its mass, m_μ. Taking m_μ as $215m_e$, where m_e is the mass of the electron, it is found that the energy of ejection of the μ-meson is 4·2 MeV.

Let m_π be the mass of a π-particle, E_μ, p_μ the kinetic energy and momentum of the μ-meson formed by its decay, and E_ν, p_ν, m_ν the corresponding quantities for the neutral particle providing the momentum balance. It then follows from the principles of the conservation of mass-energy and momentum that

$$m_\pi c^2 = m_\mu c^2 + m_\nu c^2 + E_\mu + E_\nu, \qquad \ldots\ldots(5)$$

$$p_\mu = p_\nu. \qquad \ldots\ldots(6)$$

For any assumed value of m_ν, p_ν and hence E_ν can be determined; and thence, from equation (5), m_π. In this way it can be shown that if $m_\nu = 0$, $m_\pi/m_\mu = 1{\cdot}32$. The ratio of the masses of the two types of mesons would be equal to this value if the neutral particles were photons or neutrinos. On the other hand, if the mass of the neutral particle were equal to $200\,m_e$, the value of m_π/m_μ would be 2·1. Valuable information about the neutral particle can thus be obtained from accurate determinations of the masses of the π- and μ-mesons.

3 (ii). *Mass of the π-Particles*

In approaching the problem of determining the mass of the π- and μ-particles, the difficulty is met that a single observed quantity, such as the curvature of the trajectory of a particle in a magnetic field of known intensity, depends upon both E/M, the ratio of the charge to the mass, and the velocity of the particle, v. As in the classical experiments on the electron, it is therefore necessary to make two observations which allow E/M and v to be determined separately. The value of E/M defines the mass M if the charge on the particle is assumed to be equal to that of the electron—an assumption which appears to be almost certainly correct, and for which the evidence is summarized in §4 (i). For the determination of E/M and v, a variety of methods have been employed, the most important of which may be divided into five classes.

Method (A). Residual Range and Grain-Density

The first method employed in the determination of the mass of the π-meson (Lattes *et al.* 1948) was based on the observation of (*a*) the grain-density in the tracks of individual particles in a photographic emulsion, and (*b*) the residual range of the same particles. The grain-density in a track, the number of developed grains per unit length, is proportional to the rate of loss of energy of the particle producing it. This quantity, in turn, depends only on the velocity and charge of the particle. For particles with the electronic charge the observed grain-density therefore gives a measure of the velocity in a certain range of values.

The simplest mode of application of the method is to observe the grain-density in a restricted region of the track, and thus to measure the average velocity in this

part of the trajectory, and to combine this result with the residual range. Thus equation (4) can be rewritten, for particles moving with non-relativistic velocities,

$$E = \kappa \mathcal{M}^{1-n} R^n = C \mathcal{M} v^2, \qquad \ldots\ldots(7)$$

from which it follows that

$$\mathcal{M} = k_1 R v^{-2/n}. \qquad \ldots\ldots(8)$$

The determination of the required relation between the velocity and the grain-density can be made by measurements on the tracks of long-range protons. The velocity of such a particle, for any given value of the residual range, is known from the range–energy equation—see equation (3). The grain-density in a restricted region of the trajectory can therefore be determined, together with the corresponding value of the velocity. Although measurements by this method have recently been made, its application was prevented, in the early experiments, by the limited ranges of the particles available for measurement. Any determination of the grain-density made by observations on a length of a track sufficiently restricted to ensure that the velocity was approximately constant was subject to large statistical errors as a result of the relatively small number of grains counted. To avoid these errors it was necessary to count the total number of grains in the complete track, and to modify the analysis to take account of the changing velocity of the particle along the trajectory.

We have seen that the loss of energy per unit length of path depends only on the velocity of a particle with the electronic charge: $dE/dR = f(v)$. Further, the number of grains per unit length, dN/dR, depends only on the rate of loss of energy, and we can therefore write $dN/dR = \phi(v)$. Consider two particles of mass M, m, and charge e, moving in an emulsion with velocity v. The grain-density in the two tracks will then be equal. For a given change in velocity, however, the heavier particles traverse a distance M/m times that of the lighter particles, and this result is independent of the particular values of the velocity. It follows that the total numbers of grains in the tracks of the two particles, N, n, respectively, are in the ratio of their masses: $N/n = M/m$. Since, however, the initial velocities of the two particles were the same, $R/r = M/m$, and we can write: $N/n = M/m = R/r = \mathcal{M}$.

Suppose for particles of known mass m, the relation between n and r has been determined: $n = F(r)$. It follows that, for particles of mass M, $N = \mathcal{M} F(R/\mathcal{M})$, i.e. the relation between the total number of grains in the track of a particle of any mass and the range can be computed. Alternatively, the observed relation between N and R defines the mass of the particle.

In applying this method, it is usual to plot $\log N$ against $\log R$ for the two types of particles of which the masses are to be compared. In such a diagram, points on a line inclined at 45° to either axis correspond to particles of constant velocity. Thus the range r of a proton can be written $r = \phi(v)$. It follows that $R/\mathcal{M} = \phi(v)$; but $N = \mathcal{M} F(R/\mathcal{M})$, so that $N = \mathcal{M}\psi(v)$ and $R/N = \xi(v)$, where $\phi(v)$, $\psi(v)$ and $\xi(v)$ are all related functions. For a 45° line in the ($\log N$, $\log R$) diagram

$$\xi(v) = R/N = \text{constant}.$$

The intercept of any such line with the two curves representing the results of measurements on the two types of particles gives a measure of the ratio of their masses; if the values of the range for the two intercepts are R and r, then $\mathcal{M} = M/m = R/r$ (see Figure 4).

In the early work employing this method (Lattes *et al.* 1948) it was necessary to confine the observations to the tracks of π- and μ-particles of the same event. The two related tracks of such a pair were certainly produced contemporaneously, but their relationship in time to other particles producing tracks in the emulsion, such as isolated protons, was unknown, and the effects of fading of the latent image could not therefore be estimated. In experiments on cosmic radiation at mountain altitudes, exposures of several weeks are commonly employed. As a result of fading of the latent image there was, with the types of emulsion in use at the time, a considerable difference in the grain-density of particles with the same specific ionization, according to whether the particles were recorded early or late in the exposure. The observations on the two tracks from the π- and μ-mesons of the same pair made it possible to deduce the ratio of their masses, m_π/m_μ; but the absolute values of the mass could not be determined by a comparison with the tracks of protons. A further difficulty was that in the relatively thin emulsions

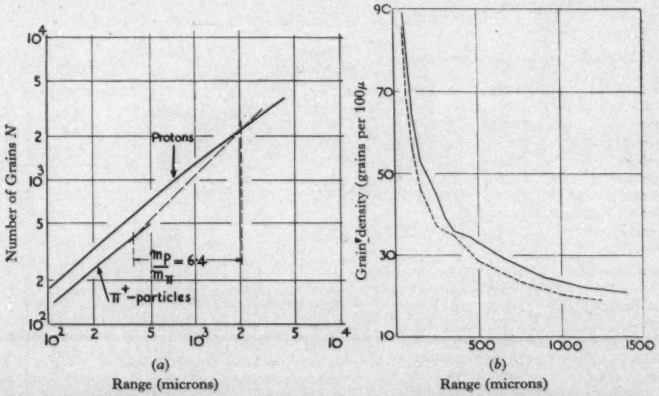


Figure 4 (*a*). Relation between log N and log R, where N is the number of grains in the track of a particle of residual range R (microns), for π^+-particles and protons produced artificially : Bowker (1950).

Figure 4 (*b*). Relation between grain-density dN/dR and range R for cosmic-ray π-mesons and μ-mesons (Brown and Fowler). Averages for five π- and five μ-meson tracks ; $m_\pi/m_\mu = 1{\cdot}35 \pm 0{\cdot}05$.

available at the time, very few long tracks, which give the most favourable conditions of measurement, were found. The value obtained for m_π/m_μ, $1{\cdot}65 \pm 0{\cdot}15$, is now known to be much too high.

Recent experiments by the method have been undertaken in the more favourable experimental conditions provided by employing new emulsions of greater thickness, and by making short exposures. A typical result obtained by Bowker in experiments on artificially generated π-mesons is shown in Figure 4(*a*). Figure 4(*b*) shows on alternative treatment of observations on π- and μ-mesons of the cosmic radiation by Brown and Fowler. The conditions provided by the former experiments are particularly favourable, for the duration of the exposure

and any fading of the latent image are negligible. The results, included in Table 1, are in good agreement with those obtained in other experiments, and prove that when applied in favourable conditions the method is capable of giving reliable results in the hands of experienced observers.

3(iii). *Method B. Scattering and Residual Range*

When a charged particle passes through a material medium, it is subject to frequent small deviations in its direction of motion as a result of Coulomb scattering. The problem was first analysed by E. J. Williams (1939). It has recently become of great importance because of its bearing on the determination of the energy of particles by the photographic method. Essentially, the observation of the average angular deviation per unit length of the trajectory, $\bar{\alpha}$, gives a measure of the quantity $p\beta$ for the particle, where p is its momentum and $\beta = v/\mathbf{c}$, the ratio of the velocity of the particle v to the velocity of light $\mathbf{c}$.

Table 1. Determinations of the Masses of π- and μ-Particles

Authors	Date	Method	Types of particles	Mass in m_e			
				m_π	m_μ	m_π/m_μ	
Lattes *et al.*	(1948)	(A) P.P.	π^+, μ^+			$1{\cdot}65 \pm 0{\cdot}15$	Cosmic-ray mesons
Goldschmidt *et al.*	(1948)	(B) P.P.	$\pi^+, \pi^-; \mu^+, \mu^-$	272 ± 12	202 ± 8		
Lattimore	(1948)	(B) P.P.	π^-	290 ± 80			
Brown *et al.*	(1950)	(A) P.P.	π^+, π^-	290 ± 20			
Barbour	(1949)	(C) P.P.	π^+, π^-, μ	270 ± 23	220 ± 26		
Franzinetti	(1950)	(C) P.P.	π^+, μ^+	281 ± 7	217 ± 4		
Camerini *et al.*	(1950)	(D) P.P.	π^+, π^-	283 ± 7			
Fretter, Retallack and Brode	(1949)	(C) W.C.	μ^+, μ^-		215 ± 2		
Gardner *et al.*	(1949)	(C) P.P.	π^-	313 ± 16		$1{\cdot}32 \pm 0{\cdot}01$	Artificially generated mesons
Bowker	(1950)	(A) P.P.	π^-	264 ± 24			
Van Rossum	(1930)	(A) P.P.	π^-	280 ± 15			
Barkas *et al.*	(1950)	(A) P.P.	π^-, μ^-	305	202		
Barkas *et al.*	(1950)	(C) P.P.	π^-	280 ± 6			
Barkas *et al.*	(1950)	(C) P.P.	π^+, μ^+	278 ± 8	212 ± 6		
Bishop *et al.*	(1950)	(E) P.P.	π^+	276 ± 6			

W.C.; P.P., indicates experiments with Wilson Chambers and Photographic Plates, respectively.

The simplest mode of application of the method would be to determine $\bar{\alpha}$, and hence $p\beta$, by observations on a restricted length of the initial part of the trajectory of a particle which subsequently reached the end of its range in the emulsion. The range–energy relation expressed by equation (4) can be employed to deduce relations between range and velocity, and hence between range and $p\beta$ for particles of any assumed mass; the mass of the particle can then be deduced from the observations.

It has recently been possible to employ this method in the case of π-mesons of relatively high energy which, by chance, reach the end of their range in the emulsion (Camerini *et al.*, unpublished). In the early experiments, however, as with Method A, the tracks available for measurement were commonly of short range, less than 1 mm. It was therefore necessary to measure the scattering over the entire length of the track and to make allowance for the variation in the

velocity of the particle. In this way Goldschmidt-Clermont, King, Muirhead and Ritson (1948) obtained values for the mass of the π- and μ-mesons which appear to be in good accord with those of later and more accurate experiments (see Table 1).

Although the method is less accurate than some others, it provides an important technical resource which has sometimes been of great value in conditions in which more refined methods cannot be brought to bear, see for example, Brown *et al.* (1949 a).

3 (iv). *Method C. Momentum and Residual Range*

(*a*) *Experiments of Barkas et al.*

The most accurate methods developed hitherto for the determination of the mass of mesons have depended on the determination of the momentum of individual particles by magnetic deflection experiments, and their residual range

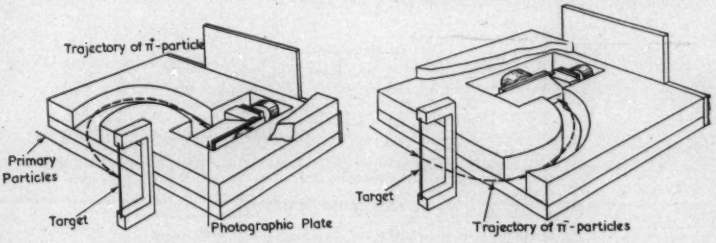


Figure 5 (*a*). Apparatus for determining the mass of π^+-particles generated artificially in the 384 in. synchro-cyclotron at Berkeley. The apparatus is placed within the vacuum tank of the cyclotron and the energetic protons or α-particles bombard the 'ribbon' target and produce π^+- and π^--particles which are emitted in all directions. The channel in the metal screens prevents particles from reaching the plates unless they have been emitted from the target in directions making angles less than about 14° with the line of motion of the bombarding particles. By employing a thin target the uncertainty in the point of origin of a particle and the radius of curvature of its trajectory in the magnetic field is reduced.

Figure 5 (*b*). Determination of the mass of π^--particles.

in a given medium. For particles of charge e, moving in a photographic emulsion, we may modify equation (3) and write:

$$E = k_2 m^{1-n} R^n,$$

where m is the mass of the particles. Further, if ρ is the radius of curvature of a particle in a magnetic field of intensity H, we may write $E = e^2(H\rho)^2/2m$, the quantities being measured in appropriate systems of units. It follows that

$$m = \left\{\frac{e^2(H\rho)^2}{2k_2 R^n}\right\}^{1/(2-n)}. \qquad \ldots\ldots(9)$$

This method has been employed by Barkas *et al.* (1950) for the determination of the mass of the π-particles generated artificially in the Berkeley synchro-cyclotron using the method illustrated in Figure 5 (*a*) and (*b*). The bombardment of the matter of the target by fast protons or α-particles leads to the creation of π^+- and π^--mesons which are emitted in all directions. Only particles which

are emitted from the thin ribbon target in directions making angles less than 14° with the line of motion of the bombarding particles are allowed to reach the recording photographic plates. The point of entry, and the direction of the initial part of the track of a given particle in the emulsion, allows its trajectory to be computed, and thence the corresponding value of $H\rho$. This observation and the observed range of the particle, when inserted in equation (9), allows the mass to be calculated. It is necessary to make corrections to allow for inhomogeneities in the magnetic field of the synchrotron. A histogram showing the results obtained by Barkas *et al.* is shown in Figure 6 (*a*). The mean value thus obtained, after suitable corrections, is $m_{\pi^-} = 280{\cdot}5 \pm 6\, m_e$.

A similar determination for the positive π-mesons, has been made by Barkas with the apparatus shown in Figure 5 (*b*), and the result thus obtained was $278 \pm 8\, m_e$. In this case observations were also possible on μ^+-particles formed by the decay of π^+-particles arrested in the target. The value thus obtained was $212 \pm 6\, m_e$ (see Figure 6 (*b*)).

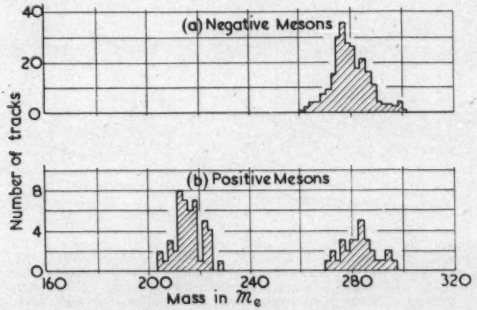


Figure 6. Histograms showing the distribution of the values of the mass of (*a*) π^-- and (*b*) π^+- and μ^+-particles, determined by Barkas *et al.*, using method C.

(*b*) *Experiments of Franzinetti, of Barbour, and of Goldschmidt-Clermont.*

Experiments similar in principle to those described above, to determine the mass of π-mesons produced by cosmic radiation, have been carried out by Franzinetti (1949), by Barbour (1949), and by Goldschmidt-Clermont (1950). The apparatus employed by Franzinetti is shown in Figure 7. The method depends upon observing the tracks of individual particles in the emulsions of two plates, placed parallel to each other and face to face, in a field of 30,000 gauss provided by an electromagnet. The two emulsions are 3 mm. apart, and attention is confined to those particles which pass across the gap at a small glancing angle to the surfaces of the emulsions, and which reach the end of their range in one of them. The tracks produced by a single particle in the two emulsions can be recognized, and the change in the direction of motion due to the magnetic field —and hence the corresponding value of $H\rho$—can thus be determined. The distribution in the values of the masses of the π- and μ-mesons obtained by Franzinetti is shown in Figure 8, and the mean values in Table 1.

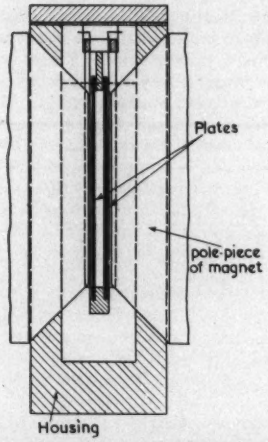


Figure 7. Apparatus employed by Franzinetti for determining the mass of particles produced by cosmic radiation in magnetic deflection experiments—Method C.

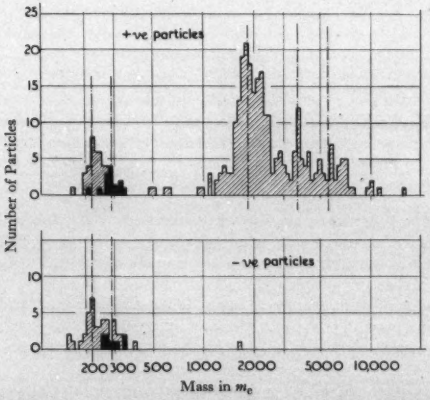


Figure 8. Distribution in the values of the mass of (*a*) +ve and (*b*) −ve particles produced at an altitude of 11,000 ft. by cosmic radiation. The results indicate that any particles of mass ~1,000 m_e are present in numbers less than 5% of the π- and μ-mesons. The black squares represent π^+- and π^-- particles definitely identified by secondary effects observed at the end of their range. (*After* Franzinetti.)

(*c*) *Experiments with Wilson chambers.*

The experiments with photographic plates are similar in principle to observations carried out over many years with Wilson chambers operated in magnetic fields. In such experiments, the curvature of a track is directly measured from the photograph, and the range of the particle deduced from the number of lead plates which it is subsequently observed to traverse. Modern forms of such an apparatus employed by Brode, and by Fretter, to which reference is made in §8 in connection with the problem of the existence of other types of mesons more

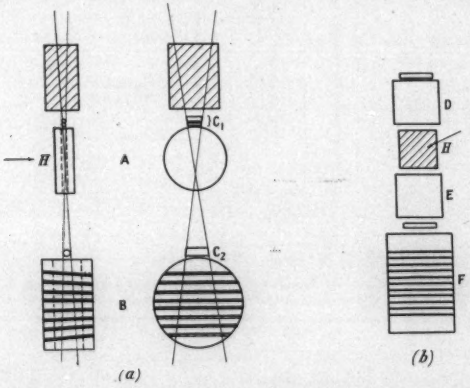


Figure 9. Apparatus of (*a*) Fretter and (*b*) Brode for the determination of the mass of particles of the cosmic radiation. In Fretter's apparatus two cloud chambers A and B are operated simultaneously by 'counter-control'—$C_1 C_2$. The upper chamber is maintained in a magnetic field of ~5,000 gauss, and serves to measure the momentum of particles which are subsequently arrested in one of the lead plates crossing the lower chamber. In Brode's apparatus, the momentum is measured by means of two Wilson chambers D and E separated by region *H* in which a magnetic field is maintained by a permanent magnet. The change in direction of motion of a particle in the magnetic field, as indicated by its track in the two chambers, therefore gives a measure of its momentum. The particle is subsequently arrested in one of the lead plates of a third Wilson chamber, F.

massive than the π-particles, are shown in Figure 9. The method has the disadvantage that the accuracy with which the range of the particles can be determined is limited by the finite thickness of the lead plates in which they are arrested, and by the fact that any secondary effects occurring at the end of the range of a particle are commonly unobserved. It has, however, been employed with success to determine the mass of the μ-particles and, in particular, has provided the only accurate information hitherto available on the mass of μ^--mesons.

3 (v). *Method D. Grain-Density and Scattering*

It has been known for many years that in the case of charged particles moving with velocity less than $0{\cdot}8c$, the simultaneous observation of the specific ionization and the momentum of the particle allows its mass to be determined. Experiments based on this principle have been made with Wilson chambers.

The specific ionization, which gives a measure of the velocity of the particles if the charge is known, has been determined by drop-counts and the momentum from the observed curvature of the trajectories of the particles due to the magnetic field.

The precision of these observations has been limited by the inaccuracy of the method of 'drop-counting'. Especially in experiments with 'counter-controlled' Wilson chambers, it has proved to be difficult to establish precisely the same degree of super-saturation in each expansion, and the drop-counts for particles of the same specific ionization are subject to uncontrolled fluctuations. In spite of these difficulties, the method has been of great importance. It allowed the μ-mesons of the penetrating component of the cosmic rays to be distinguished from electrons and protons; and it has been employed by Rochester and Butler, and by Anderson and his colleagues, in their experiments on τ-mesons described in §9.

A method very similar in principle to the above has been employed recently by Camerini *et al.* (1949), and by Fowler (1950), using photographic plates. The theory of the method has been studied independently by Goldschmidt-Clermont (1950). The velocity of the particle is determined by grain-counts, and the quantity $p\beta$ by observations on the scattering. Apart from difficulties due to fading of the latent image, which appear to be unimportant when using modern 'electron-sensitive' emulsions and short exposures, the determination of the grain-density in a track depends only on the number of grains counted and on variations in the uniformity of development and the sensitivity of the emulsion. It is therefore superior to drop-counts obtained with counter-controlled Wilson chambers. Secondly, the measurement of the quantity $p\beta$ by observations on the scattering can be made with a precision not inferior to that commonly achieved in the determination of the momentum by observations on the curvature of the tracks in Wilson chambers operated in magnetic fields.

A characteristic result obtained by this method is shown in Figure 10 in which the corresponding values, for individual tracks, of the quantities $\bar{\alpha}$, the average angular deviation of the trajectory per unit length, and g, the number of grains per unit length, have been plotted. The results all refer to particles of great range, and the variation of $\bar{\alpha}$ and g over the measured length of the track can be neglected. The lines marked π, P, D, T etc. represent the calculated distributions for π-particles, protons, deuterons and tritons, and the grouping of the experimental points round the calculated curves for π-particles and protons is particularly well marked. The distance from the line P of any observed point in such a diagram, measured parallel to the axis of the scattering parameter $\bar{\alpha}$, is proportional to the ratio of the mass of the corresponding particle to that of the proton. A characteristic mass spectrum deduced by this method—for the particles emitted in the disintegrations produced by protons and α-particles of great energy present in the cosmic radiation—is shown in Figure 11. The results prove that fast π-particles are commonly created in such disintegrations, and they allow the mass of the particles to be determined. The results are included in Table 1.

An important feature of this particular method of determining the mass of charged particles is that it can be applied to the tracks of particles of great range which escape from the recording medium. It is thus applicable in conditions in which the other four methods cannot be employed.

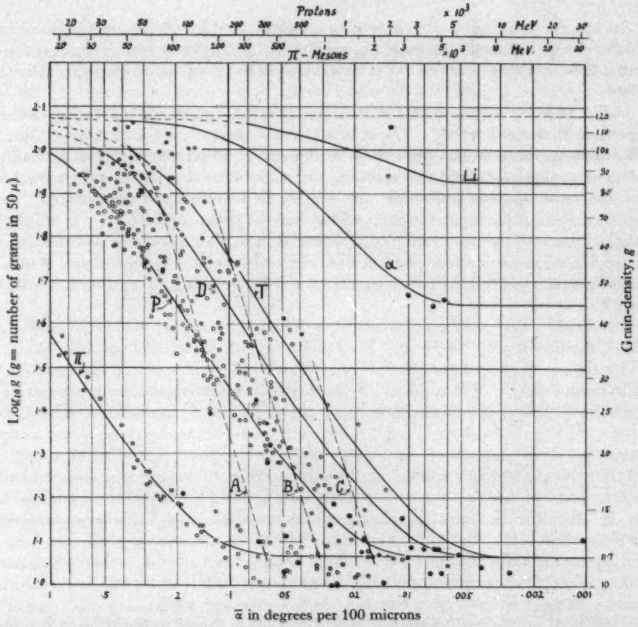


Figure 10. Distribution in the values of $\bar{\alpha}$ (the scattering parameter) and g (grain-density) for particles emitted from stars (Fowler 1950)—Method D.

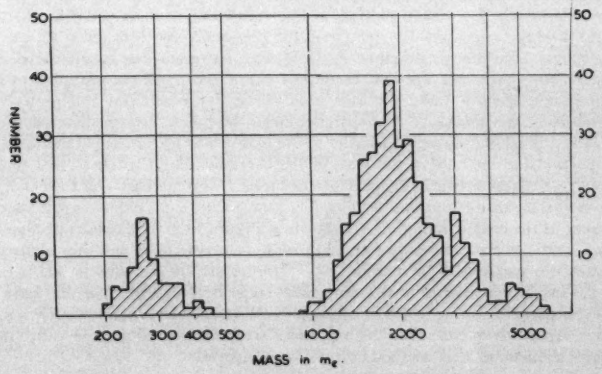


Figure 11. Distribution in the values of the mass of particles emitted from stars as determined by Method D. (Camerini *et al.* 1950.)

3(vi). *Method E. Ratio of the Ranges and of the Momenta of Protons and π-Particles*

One of the most important sources of error in the experiments of Barkas *et al.*, described under Method C—experiments which have given the most accurate values for the mass of the π-particles obtained hitherto—is that associated with an uncertainty in the precise form of the range–energy relationship for protons. In order to overcome this difficulty, Bishop, Bradner and Smith (1949) have employed a method which is independent of the absolute magnitude of the magnetic field and of a precise knowledge of the range–energy relationship.

We have seen that for particles of the same velocity, and of masses M and m respectively, we may write $M/m = R/r$. Further, since the velocities of the two particles are the same, their momenta will be proportional to their rest-masses and it follows that

$$\frac{M}{m} = \frac{R}{r} = \frac{(H\rho)_M}{(H\rho)_m},$$

a result which is independent of the form of the range–energy relation. The method therefore consists in determining that value of the ratio of $H\rho$ for protons and π-particles for which the ranges of the particles stand in the same ratio. It is carried out by means of the apparatus shown schematically in Figure 12.

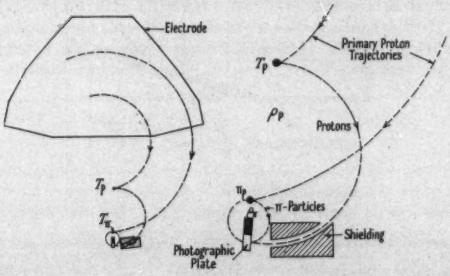


Figure 12. Apparatus of Bishop *et al.* (1949) for the determination of the mass of π^+-particles by Method E. The two ribbon targets T_p and T_π serve as sources of protons and π-mesons respectively. The two radii, ρ_p and ρ_π, are approximately in the ratio of the masses: $\rho_p/\rho_\pi = m_p/m_\pi$.

The apparatus placed in the vacuum tank of the synchro-cyclotron is provided with two ribbon targets T_p and T_π (Figure 12) designed to act as sources of protons and π-mesons respectively, and so located with respect to the photographic plate that the ratio of the energies of the two types of particles recorded is approximately equal to the ratio of the masses. In passing across the plate, there is a rapid variation in the velocity of the π-mesons and therefore in their range, and a much smaller variation in the corresponding values for protons. The observations consist in determining, for a succession of narrow elements of length along the plate, the ratios R_p/R_π and $(H\rho)_p/(H\rho)_\pi$. A characteristic set of observations thus obtained is shown in Figure 13. The interception of the best straight line through the experimental points with the line $R_p/R_\pi = (H\rho)_p/(H\rho)_\pi$ gives the ratio of the mass m_p/m_π. The value given by this method is $m_\pi = 276 \pm 6\, m_e$, and the authors state that it will be possible to obtain a more precise result by further attention to detail.

The values shown in Table 1 indicate that any difference between the masses of the π-mesons of the cosmic radiation and of those produced artificially is less than the errors of measurement. They therefore give strong support for the view that the two types of particles are identical. Further, any difference in mass between the π^+- and π^--particles is small.

3 (vii). *The Nature of the Neutral Particle emitted during μ-decay*

Since an application of the conservation laws leads to the conclusion that the neutral particle providing the momentum and energy balance during the decay of a π-meson is of small or zero rest-mass, it may be assumed to be either a γ-ray or a neutrino. Recent observations by O'Ceallaigh (1950) provide strong evidence that the first of these possibilities can be excluded, and it is reasonable to regard the neutral particle as a 'neutrino'; i.e. some form of neutral radiation of small or zero rest-mass.

Suppose the process of decay to be accompanied by the emission of a photon which recoils in the opposite direction to the μ-meson. The line of motion of the photon, and its energy $\sim$30 MEV., is defined. If, therefore, a search is made

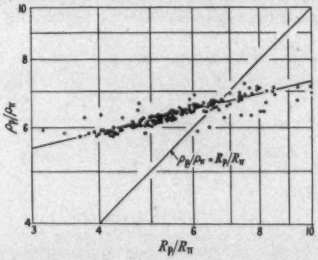


Figure 13. Observations of $\bar{R}_p/R_\pi$ for the tracks of individual π-mesons against ρ_p/ρ_π. The intercept of the 45° line with the best line through the experimental points gives a measure of m_p/m_π which is independent of a knowledge of the range–energy relationship—Method E.

of 'electron-sensitive' photographic plates in which the decay of π-mesons is recorded, it should be possible to observe pairs of electrons created by any photons recoiling from the μ-mesons. Photons of energy 30 MEV. produce pairs in photographic emulsions with a divergence of about 3° (King 1950), and the bisector of the line of motion of the two electrons defines that of the parent photon within narrow limits. Further, in favourable cases, the Coulomb scattering of the electrons enables their energies to be determined, and thence the quantum energy of the photon. Because of these features it is possible to establish whether or not an observed pair has characteristics consistent with its having arisen during the decay of a neighbouring π-meson arrested in the emulsion, or whether it is due to a γ-ray not associated with the decay.

O'Ceallaigh has scrutinized the emulsion in the neighbourhood of π-mesons stopped in plates exposed to cosmic radiation and has found no pairs in a total length of path of the recoiling neutral particles of 38 cm. The 'conversion length' of γ-radiation in the emulsion, for photons of energy equal to 30 MEV., is 6·5 cm. If the process of decay leads to the ejection of a photon, six pairs of electrons

should have been observed in the conditions of the experiments, and the probability of observing none is $<0{\cdot}005$. It may therefore be concluded that the momentum balance is provided by a neutral radiation of a different type, by some form of neutrino.

It is well known that during the β-decay of radioactive nuclei there is a disappearance of energy which has been attributed by Fermi and by Pauli to the emission of a neutral particle of low rest-mass—a 'neutrino'. Further, experiments on the masses and the β-decay of certain of the light nuclei prove that if such particles exist their rest-mass must be less than $0{\cdot}1\,m_e$. For some years it has been recognized that important evidence for the existence of the neutrino would be provided if it could be proved that the conservation of momentum, as well as that of mass-energy, required the assumption of the emission of such a neutral particle. A number of experimenters have therefore measured the recoil of the nucleus during β-decay, and have concluded that the results demonstrate the existence of a neutrino.

In the case of the decay of the π-mesons, the secondary charged particle, the μ-meson, is of small rest-mass compared with the atomic nuclei, and the energy of the recoiling neutral particle is large. The momentum and energy of the μ-meson can therefore be determined with precision. Since the possibility can now be excluded that the neutral particle providing the momentum balance is a photon, the observed values of the masses of the π- and μ-mesons and the characteristics of the μ-decay provide important additional evidence for some other form of neutral radiation—for the existence of some form of neutrino.

§4. PROPERTIES OF THE π-MESONS

4(i). *Charge of the π-Particles*

Photomicrographs of examples of the successive decay of π-particles according to the scheme $\pi \rightarrow \mu \rightarrow e$ are shown in Plate I. If charge is to be conserved in these transmutations—and if no particles with a charge much smaller than that of the electron are emitted which escape observation—the charge on the three types of particles must be equal. If the final particle formed in the process of decay is indeed an electron, a view for which the evidence is summarized in §5(iii), the charge on the π- and μ-particles must be equal to the elementary electronic charge.

It has been pointed out by Bradner that the consistency between the values of the mass deduced by different methods provides a powerful demonstration that the charge on the π-mesons is very close to the electronic charge. In particular, he has shown that the observed degree of consistency of the measurements by methods A and B, shown in Table 1, indicates that the charge in the π-particles is equal to that of the electron to within 3% (Bradner 1949).

4(ii). *Life-time of the π-Particles*

The most accurate determinations of the life-time of the π-mesons have been made at Berkeley by the method represented schematically in Figure 14. Under the impact of fast α-particles, positive and negative π-mesons are created. Those of the negative π-mesons which emerge from the target in a suitable direction and with momenta in a narrow range of values spiral in the magnetic field in the channel cut in a solid block of metal and are recorded by one or other of two

photographic plates (Figure 14). Because of the well-known 'focusing effects' associated with the spiral trajectories of particles in a magnetic field—and in the absence of any spontaneous decay of the particles—the relative numbers recorded in these two plates, per unit area, will be inversely proportional to the lengths of path in the magnetic field, i.e. as 3 to 1 for the particular experimental arrangement shown in Figure 14. If, however, the particles suffer decay in flight, the number reaching the plates will be reduced and the effect will be more marked the greater the length of path. There will therefore be departures from the simple ratio to be observed in the case of stable particles (Richardson 1948).

The method outlined above is practicable only if the life-time is of the same order of magnitude as the period of the motion, T, of the particles in the magnetic field. For particles of mass 275 m_e moving at right angles to a field of 15,000 gauss, $T \sim 10^{-8}$ sec., a value known to be close to that of the π-particles. The latest measurements by this method show that the life-time, τ_π, lies between the values 1·4 and $0{\cdot}9 \times 10^{-8}$ sec. (see, however, Martinelli and Panofsky 1950).

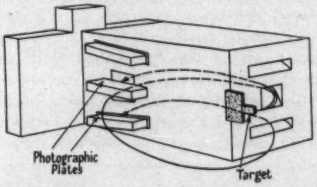


Figure 14. Apparatus of Richardson (1949) for the determination of the lifetime of the π^--particles generated artificially in the synchro-cyclotron. In the absence of decay, the numbers of particles per unit area of either photographic plate should be inversely as the length of path to the corresponding plate, owing to the effects of semi-circular focusing in one plane. The trajectory of the α-particle beam bombarding the target is not shown.

A determination of the life-time of the π-mesons of the cosmic radiation has also been made by Camerini *et al.* (1948), using a method based on the following considerations: at mountain altitudes (*c.* 11,000 ft.) the bombardment of matter by fast nucleons of the cosmic radiation stream leads to the creation of π-mesons of great energy. Some of these particles are emitted in the upward direction so that, in addition to the prominent downward flux of μ-mesons constituting the penetrating component, there is a weak upward-moving stream of mesons arising from the surface of the earth. At their point of creation in the surface layers of the earth the mesons are made up, at least predominantly, of π-particles. As their time of flight increases, the initial stream of π-particles transforms spontaneously into a stream of μ-particles. Photographic plates were therefore exposed at 2 metres above the surface of the earth. Particles of the upward moving stream which stop in the emulsion are identified by the directions of the trajectories at their points of entry into the emulsion in which they are brought to rest. Further, mesons of different types can be distinguished by the secondary processes produced at the end of their range. A study could thus be made of the rate of transformation of the π-particles into μ-particles by determining the proportion of the two types at different heights above the ground. The value thus obtained for the life-time τ_π is $0{\cdot}6 \times 10^{-8}$ sec., a result consistent with that given by the more accurate methods available with artificially produced particles.

The short life-time of the π-particles allows us to account for the failure to observe them in experiments with expansion chambers and counters. It will be shown in a later paragraph that most of the π-mesons of the cosmic radiation are emitted during nuclear explosions with a kinetic energy greater than 100 MeV. When moving in the atmosphere, the rate of loss of energy due to ionization is so slow that only a minute fraction of the particles are brought to rest before decaying into μ-mesons. It follows that there is only a very small chance of observing the spontaneous decay of a π^+-meson in the gas of a Wilson chamber.

4 (iii). *Disintegrations produced by π^--Particles*

The original observations of nuclear disintegrations produced by mesons arrested in photographic plates were interpreted as due to μ^--particles captured by silver or bromine nuclei. It is now established that at least most of them, and possibly all, are in fact due to π^--particles. The evidence rests on the following observations:

The mesons observed in photographic plates exposed at high altitudes are composed of π- and μ-particles in proportions which depend upon the amount of matter in the immediate vicinity of the emulsions. This is due to the fact, discussed in more detail in a later paragraph, that the great majority of the mesons in the atmosphere are μ-particles produced by the decay of π-particles. Most of the π-particles, the primary products of nuclear explosions, are created with a kinetic energy less than 10^9 ev.; and because of their short life-time, they commonly travel distances of less than 50 metres before decaying 'in flight'. In exposures at a height of approximately 10,000 ft., the downward stream of μ^+- and μ^--particles thus formed, most of which are moving in directions inclined at less than 40° to the vertical, can be distinguished by the inclination of the tracks at the points of entry into that emulsion in which the particles are brought to rest. On the other hand, the majority of the π-particles are locally generated. Their directions of motion depend on the location of the neighbouring matter, and are commonly distributed nearly isotropically. Further, in plates exposed in isolation from neighbouring matter, the π-particles are found to be present in numbers less than 10% of the μ-mesons, but their numbers can be increased by concentrating lead blocks round the plates during the exposure.

In a series of exposures made at a wide range of altitudes, and with varying amounts of local matter of high atomic number, it has been found that the number of π^+-particles recorded, per cm^3 of emulsion, is nearly equal to that of the mesons which produce nuclear disintegrations, σ-mesons, and that there is no significant variation in the proportion of the two types in a wide range of experimental conditions. This observation can be easily explained if the π^+- and π^--particles are created in approximately equal numbers in the bombardment of matter of high atomic number by cosmic radiation, and if the π^--particles alone produce nuclear disintegrations with the emission of protons and other charged particles when captured by the nuclei present in the emulsion.

On the other hand, if an appreciable proportion of the μ-particles also produced disintegrations, the number of σ-mesons would be correspondingly increased, and the ratio N_σ/N_π would vary with the experimental conditions and, in particular, with the amount of local matter round the plates. The absence of such an effect gives strong support for the view that very few or none of the μ^--particles produce disintegrations with the emission of protons or α-particles.

The experiments at Berkeley on negative π-particles produced artificially give very strong support for this view. Although μ^--particles are sometimes recorded in these experiments, their points of origin are usually unknown, for it is probable that π-particles alone are produced directly in nucleon–nucleon collisions, and that μ-particles arise only by the decay of the heavier particles. The identification of the μ^--particles is therefore difficult. On the other hand, the π^--particles can be identified with certainty. It is thus possible to determine the relative frequency with which π^--particles, when brought to rest in the emulsion, produce disintegrations with the emission of different numbers of protons, α-particles, etc. The distribution thus obtained (Adelman *et al.* 1949) is shown in Figure 15 (*b*). The corresponding distribution for the cosmic-ray mesons is given in Figure 15 (*a*).

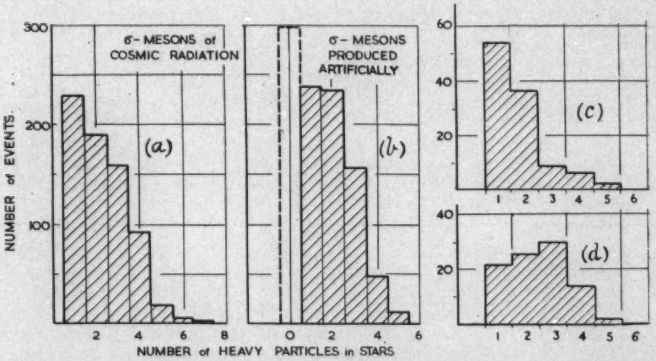


Figure 15. 'Prong distributions' for the disintegrations produced by π^--particles in photograph c emulsions: (*a*) for cosmic-ray particles; (*b*) for artificial particles; (*c*) and (*d*) show similar results, attributed to heavy and light elements in the emulsion respectively.

The two distributions shown in Figure 8 are indistinguishable within the limits corresponding to statistical fluctuations. This result can be interpreted either as indicating that the π^--particles alone produce disintegrations, or that the disintegrations produced by μ^--particles are similar in type to those produced by π^--particles. Support for the former assumption was provided by experiments on determinations of the mass of the σ-mesons by Goldschmidt–Clermont *et al.* (1948) and by Lattimore (1948), using the scattering method. The mean values obtained, $m_\sigma = 272 \pm 12$ and $m_\sigma = 290 \pm 80\ m_e$, were in accord with the assumption that the σ-mesons are π^--particles.

Recent experiments by Franzinetti, in which the masses of cosmic-ray mesons have been determined by magnetic deflection experiments, show that less than 5% of the μ^--particles stopped in the emulsion produce nuclear disintegrations in which heavy charged particles with a range greater than 5 μ are emitted.

Whilst this body of evidence is not conclusive, it shows that at most only a small proportion of the μ^--particles produce observable disintegrations in photographic emulsions, and it is consistent with the view that none of them do so. At least the great majority of the σ-mesons are therefore to be identified as π^--particles. The Berkeley experiments also prove that about 27% of the π^--particles fail to produce recognizable disintegrations. In these particular cases it may be assumed that nuclear capture followed by disintegration does occur, but that fast neutrons are liberated which escape observation and that, if they are accompanied by charged particles, the latter are of very short range. Further, the recent experiments by Panofsky *et al.* (1950) shows that in the interaction of a π^--particle with hydrogen, the most common element in the emulsion, no visible star is produced (see, however, § 8 (vi)).

Experiments with artificially produced π^--particles (Bradner 1950) show that when stopped in photographic emulsion they never, or only very rarely, decay with the emission of μ^--particles, the process by which the majority of the μ^--mesons are created in the atmosphere. We must therefore conclude that, unlike the μ^--particles, the π^--particles commonly interact with nuclei and produce disintegrations before they have had time to decay; and this is in spite of the fact that the life-time of the π^--particles is only about one two-hundredth part of that of the μ-particles.

It is reasonable to assume that the process of atomic capture of a π^--particle is similar to that of a μ^--particle. The above result therefore indicates that a π^--particle has a much stronger interaction with nucleons than a μ^--particle, that, having reached its state of lowest energy, the π^--particle interacts with the nucleus of a light element such as carbon in a time short compared with 10^{-8} sec.

4 (iv). *Capture of π^--Particles by Particular Elements*

The process of nuclear capture of π^--particles has been studied by Heidmann and Leprince-Ringuet (1948) and by Perkins (1949), who have suggested that the first interaction takes place between a π^--particle and a pair of nucleons. By studying the fate of artificially produced π^--particles brought to rest in a 'sandwich' made up of successive layers of pure gelatine and normal photographic emulsion, Menon *et al.* (1950) have shown that it is possible to distinguish some of the characteristics of the disintegrations of light elements on the one hand—carbon, nitrogen or oxygen—and of silver or bromine on the other. Characteristic nuclear disintegrations of the light and heavy nuclei distinguished by these methods are shown in Plates II and III.

In studying the disintegration of light nuclei by π^--particles it is usually difficult to distinguish the particular element involved. The disintegration is commonly accompanied by the emission of neutrons of which the energy and directions of motion cannot be determined. In certain rare instances, however, in which a single neutron is emitted in addition to the charged particles, it appears to be possible to determine the momentum of the uncharged particle by an application of the principle of the conservation of momentum, and thus to examine the total energy distributed among the emitted nucleons. A particular example of this type is the reaction ${}^{12}_{6}\mathrm{C}+\pi\rightarrow{}^{4}_{2}\mathrm{He}+{}^{4}_{2}\mathrm{He}+{}^{3}_{1}\mathrm{H}+{}^{1}_{0}\mathrm{n}$ (Menon *et al.* 1950). Such a study would appear to be of great importance for the extension of our knowledge of the process of nuclear capture by π^--particles. It is the

opinion of some authors that in the initial process of capture of π^--particles the energy liberated by the disappearance of the rest-mass is commonly shared amongst a considerable number of nucleons; amongst an α-particle group for example. This may lead to the emission of one or more fast nucleons and the formation of an excited nucleus which subsequently 'evaporates'. For a detailed discussion of these questions see Menon *et al.* (1950), Tamor (1950), Adelman and Jones (1949), and Cheston and Goldfarb (1950).

4 (v). *Spin of the π^--Particles*

It has been suggested by Wentzel (1949) that the π^--mesons created in the interaction with matter of a directed beam of fast nucleons might be 'polarized': that there might be some degree of orientation of an axis of spin of the particles, with respect to the line of motion of the 'primary' nucleons which produced them. Further, it is possible that if the particles are arrested in solid materials, the Coulomb forces brought into play in the atomic encounters which cause the particles to lose kinetic energy would not completely destroy the initial polarization. Such a polarization, if it exists, might be made manifest by privileged directions of emission of the μ^+-particles, created in the spontaneous decay of π^+-particles when at rest, with respect to the line of motion of the 'primary' nucleons.

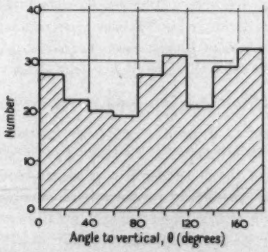


Figure 16. Distribution—relative to the vertical—of the directions of emission of μ^+-particles produced by the decay of π^+-particles of the cosmic radiation arrested in photographic emulsions.

At mountain altitudes (11,000 ft.) the cosmic-ray stream contains a large proportion of protons and neutrons, most of which move downwards in directions inclined at less than 40° to the vertical. In traversing photographic plates these particles produce nuclear explosions some of which are accompanied by the emission of π^+-mesons, and some of these particles stop in the emulsion. An examination of the directions of motion of the μ^+-particles emitted in the decay of π^+-particles can thus be made, and a typical result is shown in Figure 16. It will be seen that the particles are emitted at all directions to the vertical and any departures from an isotropic distribution are small.

In the above experiment we are observing the decay of π-mesons which are emitted from nuclear explosions with a wide range of values of the kinetic energy. Further, in many cases the particles may have been produced as one of a number

of mesons created in a single nuclear event—as one of a penetrating shower—see §7 (ii). In these circumstances the absence of an observed anisotropy is indecisive and no conclusions regarding the nature of the spin of the π-particles can be drawn from the experiments.

Similar conclusions have been drawn from preliminary experiments made with artificially produced π^+-particles by Richman *et al* (1950).

Decisive evidence in support of the view commonly held that the π- mesons have spin zero or unity is provided by the experimental evidence of Panofsky *et al.* (1950) on the capture of π^--particles by protons and of the mode of disintegration of the resulting neutral mesons (see §8 (ii)).

4 (vi). *β-Decay of the π-Particles*

The properties of the π-particles set out in the preceding sections are closely similar to those attributed to Yukawa particles, except that the former commonly suffer μ-decay instead β-decay. Observations have been made at Berkeley to determine whether the spontaneous transformation of a π-meson invariably leads to the emission of a μ-meson, or whether it sometimes undergoes β-decay. The proportion of the π^+-particles, identified by their mass, which stop in the emulsion and emit a μ-meson has been determined by means of an apparatus of the type shown in Figure 5 (*b*).

The experimental results indicate that at least 95% of the π^+-particles undergo μ-decay, and not more than 5% direct β-decay. If therefore we assume that the π-meson can decay in either of two modes, we can set an upper limit to the decay constant for the process, which leads to the emission of an electron, and an equivalent upper limit to the effective life-time. The result thus obtained is as follows:

$$\tau_\pi(\beta) > 20 \times 10^{-8} = 2 \times 10^{-7} \text{ sec.}$$

This value is very close to that postulated by Yukawa for the life-time, against β-decay, of his heavy quanta.

§ 5. PROPERTIES OF THE μ-MESONS

5 (i). *Mass of the μ-Particles*

The most accurate determinations of the mass of the μ-mesons have been made by measurements on the curvature of the trajectory of the particles in Wilson chambers operated in magnetic fields, together with the corresponding values of the residual range in lead plates—method C (Fretter 1946, Retallack and Brode 1949, Brode 1949). The essential features of the methods are illustrated in Figure 9 and the results summarized in Table 1.

The mass of the particles has also been determined by observations on the tracks in photographic plates by grain-counting and by the scattering method (Goldschmidt-Clermont *et al.* 1948), but the results are not so reliable as those obtained in the best experiments with the Wilson chamber. Recently, determinations have also been made by magnetic deflection experiments, detecting the particles with photographic plates, see method C, §3 (iv). The observations, by this method, on the μ^+-particles produced by the decay of artificially generated π^+-particles appear very promising—see Table 1 and Figure 6.

5 (ii). *The Decay of the μ-Meson*

Reference has already been made to the original observation by Williams and Roberts (1940) which demonstrated that the μ-meson sometimes decays with the emission of a charged particle of energy ~50 MeV. and low rest-mass, and to the support which was thus given for the view that the particle transforms into an electron together with a neutral particle also of low or zero rest-mass —a neutrino or a photon. The evidence that the emitted charged particle is an electron, summarized in a later paragraph, is now very strong. Although its nature remains to be finally established, it will be convenient in what follows to refer to it as an electron.

No further successful observations with Wilson chambers of the decay of mesons were made for a number of years. In 1947, however, Anderson obtained two examples in both of which the energy of the electron appeared to be of the order of 25 MeV. Later, other observers, employing a variety of methods, obtained indications that the energy of the electron is not unique (Thompson 1948, Fowler *et al.* 1948, Zar *et al.* 1948 and Steinberger 1948, 1949). Steinberger deduced the energy distribution of the electrons by observations on the ranges of the particles in solid materials. Leighton, Anderson and Seriff (1949) on the other hand determined their momenta from the curvature in the tracks in a Wilson chamber operated in a magnetic field at sea level. Finally, with the development of the electron-sensitive emulsions, it also became possible to record the tracks of the decay particles and, in favourable cases, to determine their energies by the scattering method. In this way it was shown by Brown *et al.* (1949 a) that the values of the energy of ten of the particles were distributed in the interval from 10 to 50 MeV.

It is now certain that the energy of the electron emitted during the decay of the μ-meson is not constant but may have any value distributed in a range up to a maximum of about 55 MeV. In the decay of a meson into an electron and some form of neutral radiation, a simple application of the principles of the conservation of energy and momentum enables the maximum permissible value for the energy to be calculated. The value quoted above, 55 MeV., corresponds to the maximum energy of the electron when the rest-mass of the meson is 215 m_e, the neutral particle, or particles, providing the momentum balance being of zero rest-mass. The best determination of the form of the energy 'spectrum' is that of Leighton *et al.* (1949), which is reproduced in Figure 17 (*a*). The results indicate that there is a maximum in the distribution at about 35–40 MeV., and a finite probability of the decay particle being ejected with the maximum possible energy. These results are consistent with those of Steinberger.

A second determination of the energy spectrum has been made by Davies *et al.* (1949), applying the scattering method to a study of the tracks of the decay particles in photographic emulsions. The result obtained is shown in Figure 17 (*b*). It is a feature of this method that a homogeneous group of particles appears as an asymmetrical peak in the energy distribution, with a 'tail' extending to higher energies. The presence of a small proportion of particles with an apparent energy greater than that theoretically possible, shown in Figure 17 (*b*), is therefore not unexpected. The authors show that the mean mass of the mesons which were the parents of these particular tracks is $204 \pm 19\, m_e$. Most of them must therefore be attributed to μ-mesons, and not to the presence of a small proportion of heavier

particles with a similar mode of decay. It has been pointed out by O'Ceallaigh, however, that some of them may be due to the direct β-decay of π^+-particles.

In four cases the authors have been able to determine the energy of the electrons arising from μ-mesons which were themselves produced in the emulsion by the decay of π-particles. The values thus obtained fall in the range of energies from 10 to 50 MeV. and give additional support for the view that the μ-mesons are identical with the mesons of the penetrating component of the cosmic radiation. Examples of the successive decay π–μ–e are shown in Plate I.

It has been remarked by Rossi (1949) that, in order to study the decay products of free μ-mesons, attention should be confined to the positive particles, for the negative particles are believed to fall into Bohr orbits round atomic nuclei when arrested in solid materials. Strong evidence for the existence of the atomic

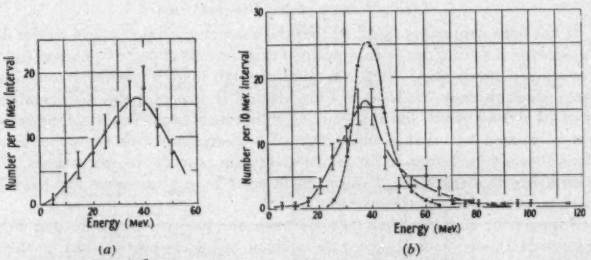


Figure 17 (*a*). Distribution in energy of the electrons produced by decay of μ-particles. Observations of Leighton *et al.* (1949).

Figure 17 (*b*). Distribution in energy of the electrons produced by the decay of μ-particles. Observations of Davies *et al.* (1950); the dotted curve corresponds to the distribution to be expected for a homogeneous group of electrons of energy 40 MeV.

capture of mesons has been provided by experiments of Cosyns *et al.* (1949). These authors have studied the distribution in energy of slow electrons, in the interval from 10 to 50 kev., which originate at the end of the range of about 10% of the μ-mesons arrested in photographic plates exposed to cosmic rays. The authors show that the distribution in energy of these electrons can be accounted for in terms of Auger transitions accompanying the atomic capture of μ^--particles by silver and bromine. These transitions result in the ejection of atomic electrons as the mesons fall to states of low energy round the nucleus (see also Chang 1949). The μ^--particles only decay, however, when captured by light elements. While the process of atomic capture must occur in these elements also, the loss of energy of the decay electron in escaping against the Coulomb attraction of the nucleus will be small—not more than approximately 1,000 ev. It is therefore reasonable to suppose that the effect on the form of the energy distribution of the decay electrons is small, and that the results by the photographic method, some of which correspond to the decay of μ^--particles, are not seriously in error from this cause.

Since the energy of the electron is not unique, it must be assumed that at least two neutral particles are emitted during its transmutation. Both the distributions shown in Figure 17 indicate that there is a maximum at about 40 MeV.,

and that the mean energy of the electron is about 35 MeV. This value is about one-third the total energy liberated by the disappearance of the rest-mass of the μ-meson. This result suggests that two neutral particles of small rest-mass are created, the average energy of each of the three particles being one-third of the available energy.

The form of the energy 'spectrum' of the charged and uncharged decay particles is of great importance for determining the nature of the forces between them. The expected distributions for various types of forces have been calculated by Tiomno, Wheeler and Rau (1949). Some of them, notably the vector and pseudo-scalar varieties, appear to be in satisfactory agreement with the observed distributions. It is important to increase the accuracy and statistical weight of the observations in order that a more precise comparison between theory and experiment may be possible (see also Michele 1950).

5 (iii). *Nature of the Decay Particles*

It has been commonly assumed that the charged particle formed by the decay of a μ-meson is an electron. In support of this view Hincko and Pontecorvo (1949) have studied the passage of the particles through matter and the rate of loss of energy through *bremsstrahlung*. This process is less probable the greater the rest-mass of the particle, and observations of the rate of loss of energy of the particles when compared with that of an electron with an energy within the same range of values allows conclusions to be drawn about the ratio of the rest-masses. The experiments show that the decay particle must have a rest-mass less than $2\,m_e$, where m_e is the mass of the electron.

Support for this view has recently been obtained by Camerini and Fowler (unpublished), who have observed the collision of a decay particle with an electron in a photographic plate. The energies of the initial particle and of the two components producing the forked track resulting from the collision can be determined by the scattering method. The analysis of the event indicates that mass of the decay particles is $(3 \pm 2)m_e$ or $(1{\cdot}2 \pm 0{\cdot}5)m_e$, according to which arm of the fork is assumed to be that of the recoiling electron.

This evidence therefore proves that if the decay particle is of a type of which the existence is already established, it is an electron, and that, in any case, its rest-mass is less than $2\,m_e$.

We have seen that the distribution in energy of the charged particles liberated by the decay of the μ-meson indicates that at least two neutral particles are also created. There is very strong evidence that this neutral radiation is not made up of photons. Thus Hincks and Pontecorvo (1948) have searched for coincidences due to the decay electron from a μ-meson on the one hand, and a discharge in other counters resulting from the materialization of any accompanying γ-radiation on the other. They have found that any such quanta, if they occur at all, appear much less frequently than the charged decay particles, and similar results have been obtained by Sard *et al* (1948).

In view of these results, and of the fact that the experimental evidence suggests that the maximum energy of the decay particles is of the order of 55 MeV., it is generally assumed that the neutral radiation must consist of particles of small or zero rest-mass, of some form of 'neutrino'. The decay of the μ-meson is therefore commonly represented by the equation

$$\mu \rightarrow e + \nu + \nu,$$

where ν represents a neutrino. This decay scheme, and that for the π-mesons represented by the equation $\pi \rightarrow \mu + \nu$, are consistent with the assumption that the μ-meson is of spin $\frac{1}{2}$ and the π-meson of spin 0 or 1 (Marshak 1949, Serber 1949). It remains to be established, however, whether the three neutral particles represented in the above equations are of the same type, and whether or not they are identical with the neutrino of the nuclear β-decay.

5 (iv). *Life-time of the μ-Particles*

A modern form of apparatus to determine the life-time of μ-mesons of the cosmic radiation, due to Valley and Rossi (1949), is illustrated in Figure 18. 'Delayed coincidence' counters, similar in their mode of operation to those illustrated in Figure 1, are employed in conjunction with an expansion chamber operated in a magnetic field. The sign of the charge of a meson arrested in the

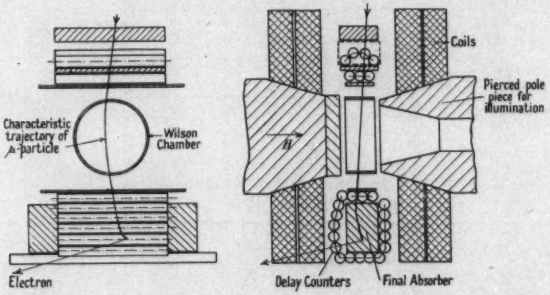


Figure 18. Apparatus for the determination of the life-time of μ-particles (Valley and Rossi 1948). The sign of the charge of the incoming particle is determined by the sense of its deflection in the Wilson chamber operated in a magnetic field.

final absorber is deduced from the sense of its deflection in the magnetic field. The best value for the mean life-time obtained by this method is $2{\cdot}15 \times 10^{-6}$ second.

Preliminary reports have recently appeared (Alvarez 1950) of a determination of the life-time of μ^+-particles produced artificially. When energetic protons or γ-rays fall on matter, π^+- and π^--mesons are created (see §6 (iv)) which can be arrested in neighbouring absorbers. The π^+-particles then decay to give μ^+-particles which in turn produce decay electrons. The total time between the instant of formation of a π^+-particle and the arrest of the resulting μ^+-particle is of the order of 10^{-8} second, a period short compared with the life-time of the μ-particles. Advantage can therefore be taken of the characteristics of the synchro-cyclotron of producing short pulses of protons, detecting the decay electrons with scintillation counters. A pulse leads to the sudden creation of a large number of π^+-particles, and the distribution in the times of delay in the emission of the decay electrons arising from the μ-mesons is determined. It is thus possible to make observations of great statistical weight, and it appears probable that the method will yield very accurate values for the life-time of the μ-mesons. A characteristic example of the observed distribution of the delay times obtained by this method is shown in Figure 19; see also Steinberger *et al.* (1950).

5 (v) *Nuclear Interactions of μ^--Particles*

The important bearing which the study of the behaviour of μ^--particles, when stopped in materials of different atomic number, has had on the historical development of our knowledge of mesons has already been stressed in an earlier paragraph. In materials of low atomic number the process of spontaneous decay predominates, and the delayed coincidence experiments indicate that the life-time of the particles is $2{\cdot}15 \times 10^{-6}$ sec., whereas in heavy elements no delayed coincidences are observed. These effects were attributed to nuclear capture, and there is now decisive evidence that this explanation is correct. No decay electrons are observed, for example, from μ^--particles stopped in silver bromide. At the time

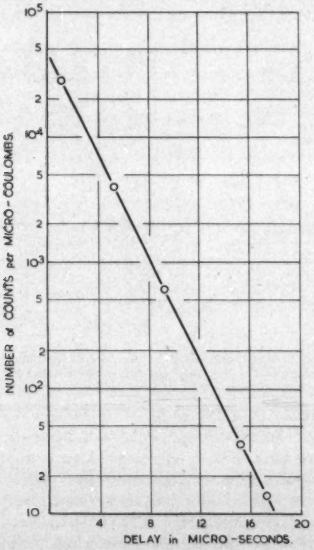
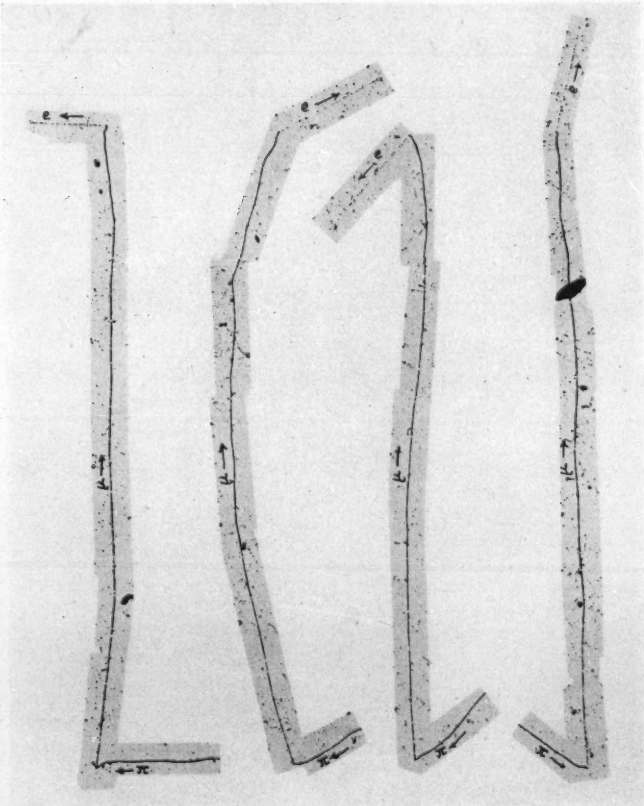


Figure 19. Observations on the delay in the instants of decay of artificially produced μ^+-particles (Alvarez *et al.* (1950)).

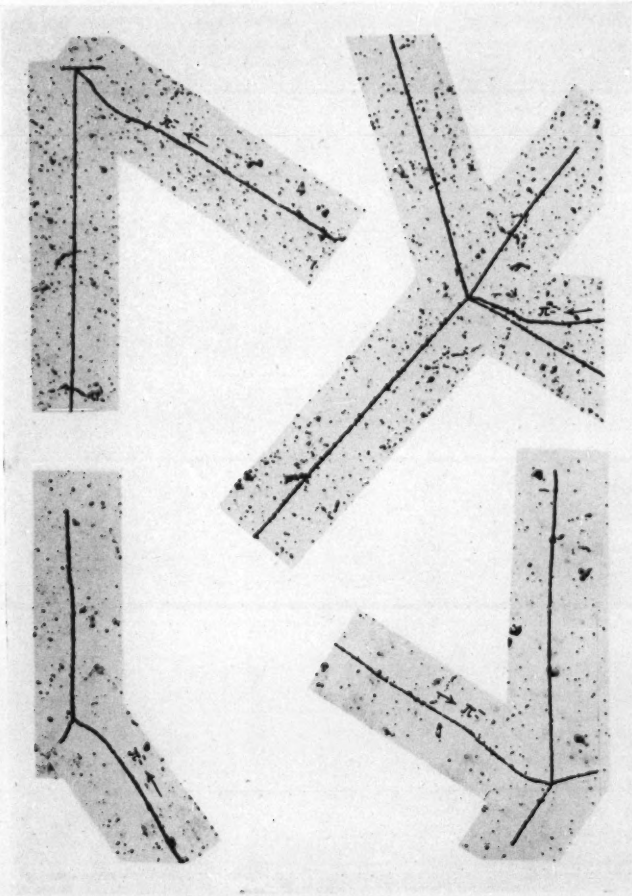
of the first observations, however, an alternative possibility was not excluded: that the absence of delayed coincidences was to be attributed to an acceleration of the rate of decay of the meson in the strong field of the nucleus (Valley 1945). In this case the μ-particles would emit electrons, but with delay times less than those characteristic of a mean life of the order of 2 μsec. With the resolving time of the counters then available, such electrons would appear in coincidence with the meson, i.e. without detectable delays. In these circumstances the assumption of an accelerated decay was eliminated by ingenious experiments of Ticho and Schein (1948) and Valley and Rossi (1948) on the behaviour of μ^--particles arrested in



Four examples of the successive decay $\pi \rightarrow \mu \rightarrow e$. The photo-micrographs display the increase in the grain-density of the tracks of the μ-mesons as the particles approach the end of their range. The sparseness of the grains in the tracks of the electrons may be clearly seen.

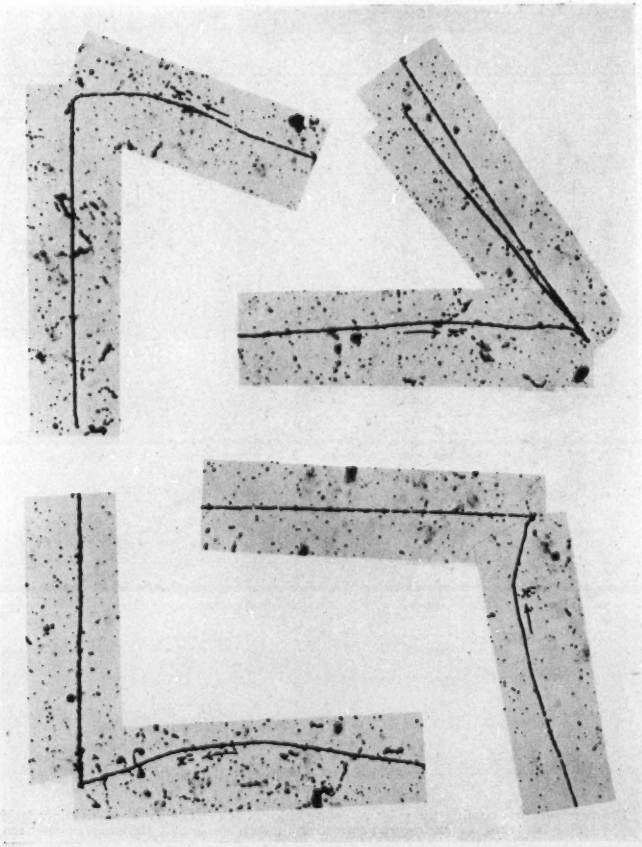
PLATE I.

To face page 384



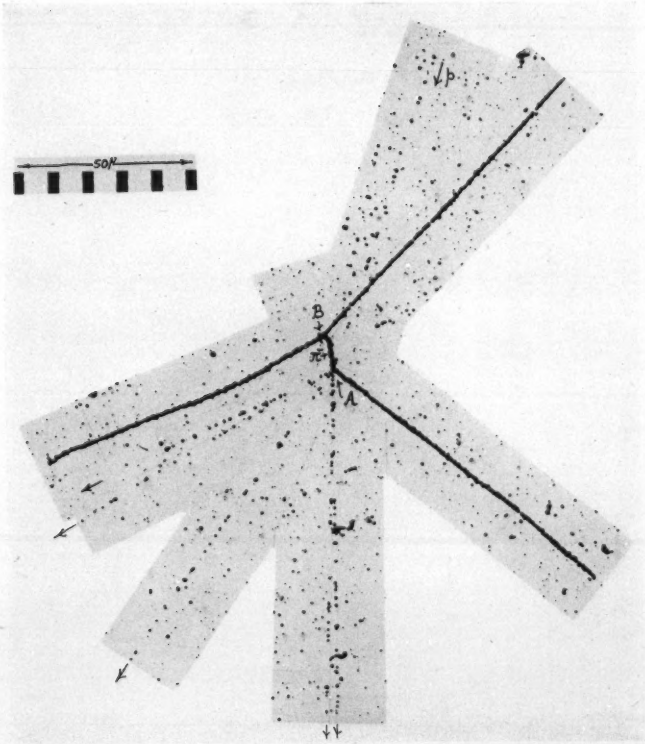
Mosaics of photo-micrographs of the disintegrations of the nuclei of the light elements in the emulsion, carbon, oxygen or nitrogen, following the capture of a π^--particle; see Menon *et al.* (1950).

PLATE II.



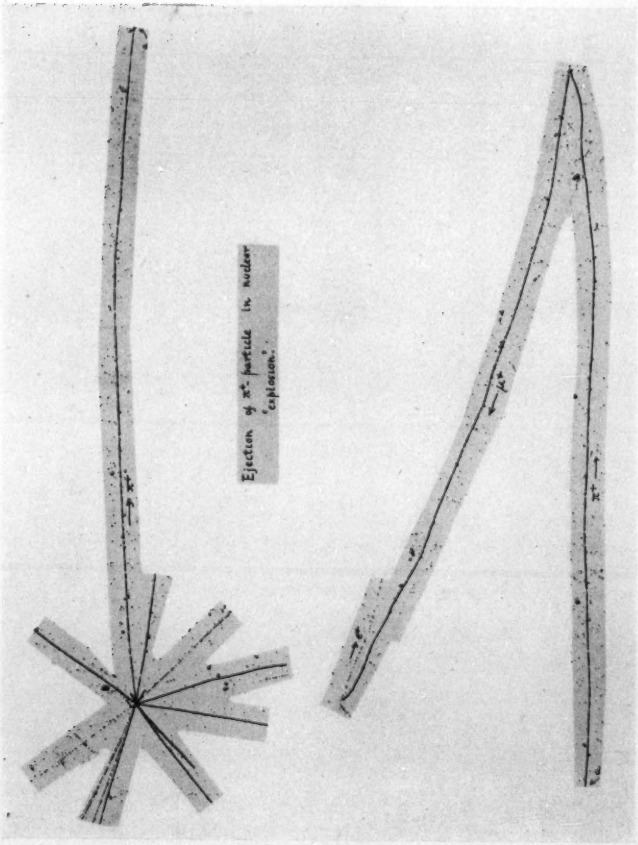
Disintegrations of heavy elements by the capture of π^--particles.

PLATE III.



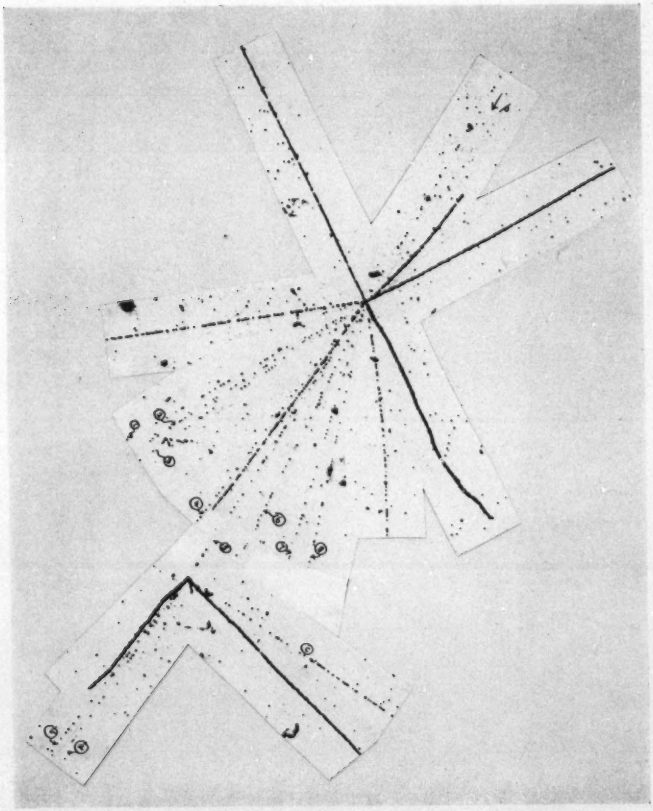
Disintegration produced, at the point A, by a particle, track p, moving at relativistic velocities. Five fast ' shower ' particles and a negative π-meson are ejected; the latter, of short range, comes to rest at the point B and produces a second disintegration. In addition, a single heavily ionizing particle is emitted of which the nature cannot be established. It was probable a proton.

PLATE IV.



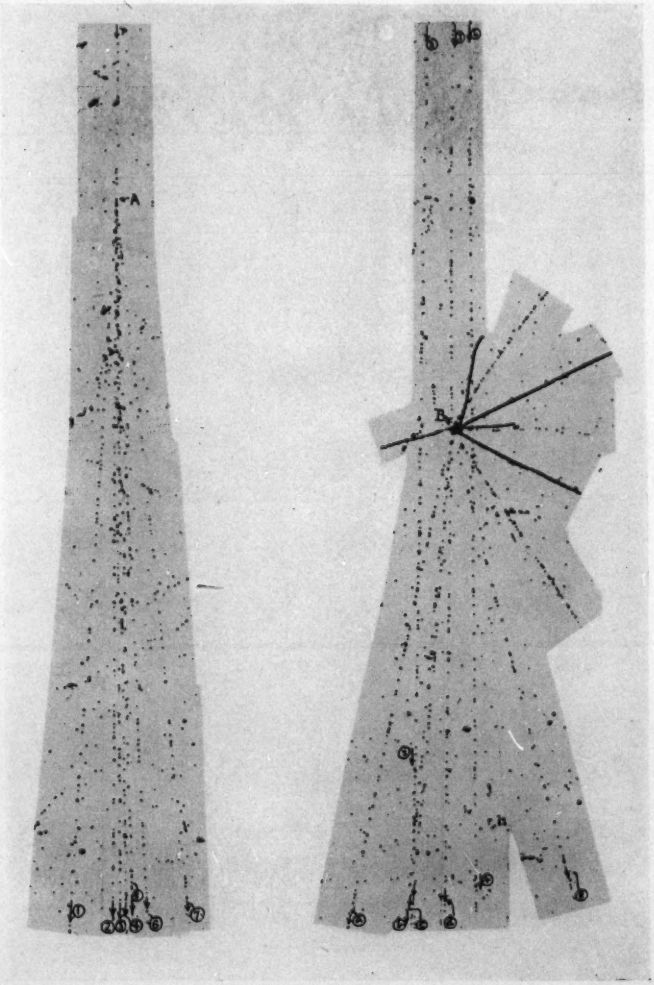
Mosaic of photo-micrographs of a nuclear explosion accompanied by the ejection of a π^+-particle. The track of the π^+-particle is given in two parts which should join at the point 'a'. The π^+-particle shows the transmutation $\pi \rightarrow \mu \rightarrow e$.

PLATE V.



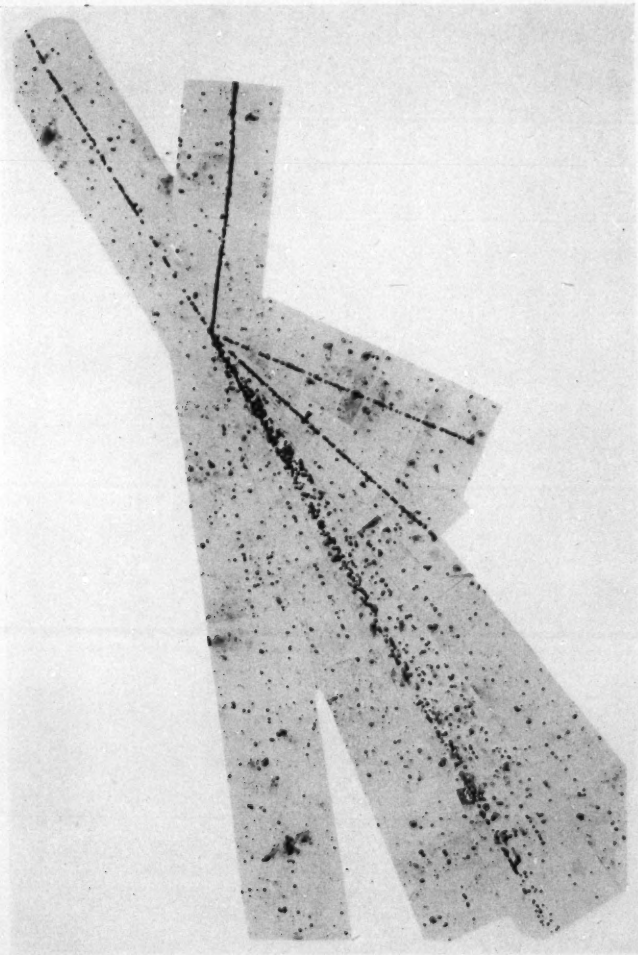
Characteristic ' double-star '. One of the ' shower ' particles emerging from a nuclear explosion produces a second disintegration.

PLATE VI.



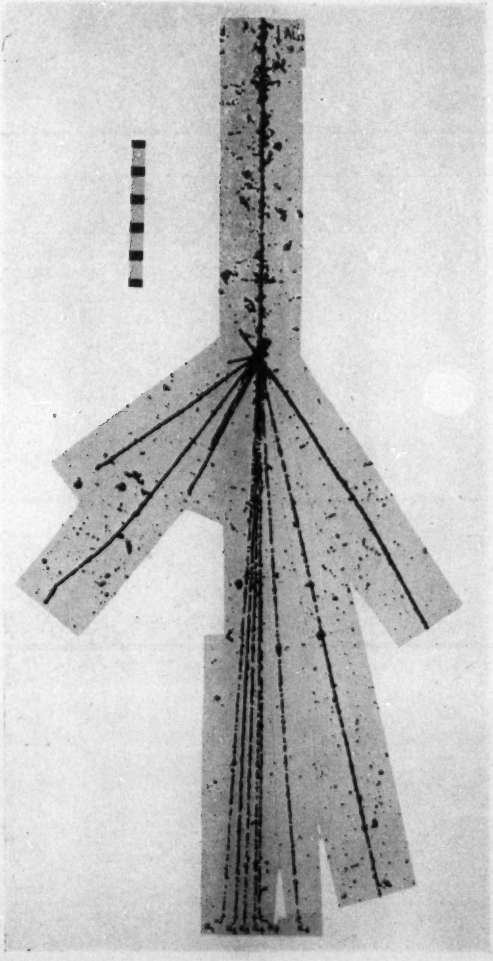
A fast particle, track p, makes a nuclear collision at the point A, as a result of which seven ' shower ' particles are produced. No. 3 of the shower particles produces a second ' shower ' at the point B. The event at point A may be due to a proton–proton collision.

PLATE VII.



An α-particle of great energy, of which the track enters the field-of-view, top left, interacts with a nucleus and produces thirty-five shower particles, most of which are in the main 'jet'.

PLATE VIII.



A nucleus of aluminium ($Z=13\pm1$), makes a nuclear collision from which a jet of six α-particles emerges with nearly equal velocities. They must be attributed to the disintegration of the original nucleus.

PLATE IX.

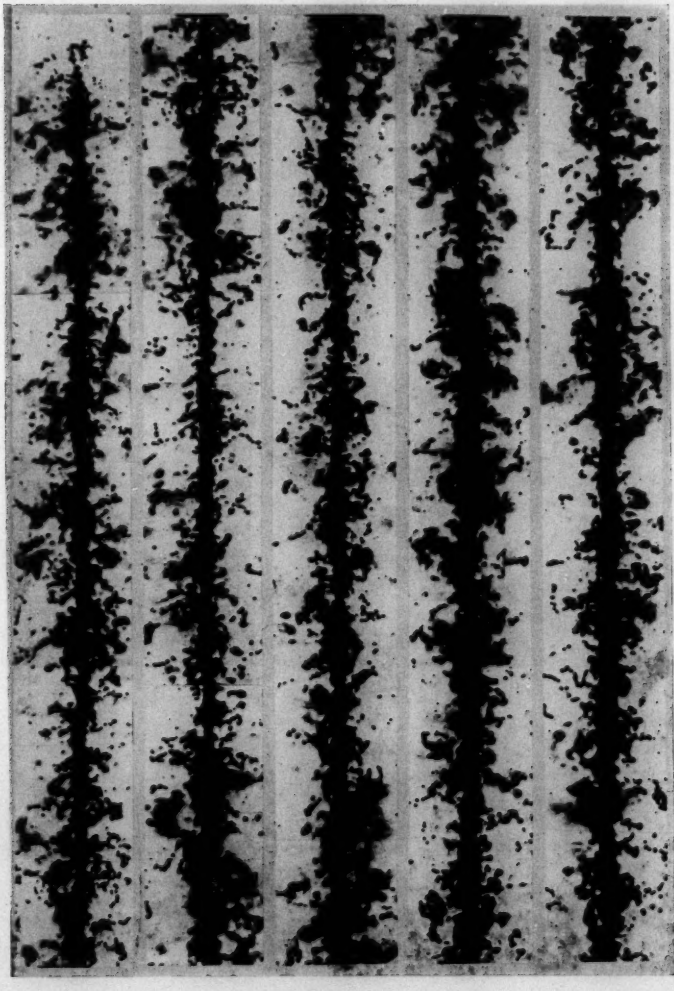


PLATE X.

Plates X and XI. Successive overlapping sections in the track of an iron nucleus ($Z = 26 \pm 2$) which between sections seven and eight is due to the particle crossing an air-gap between two and other features associated with the capture

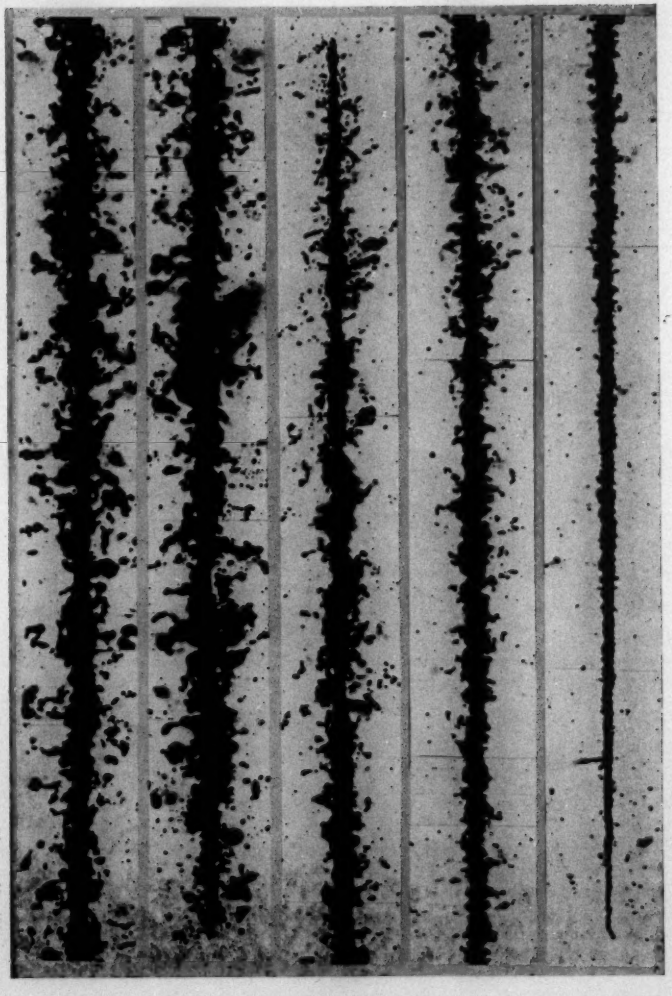
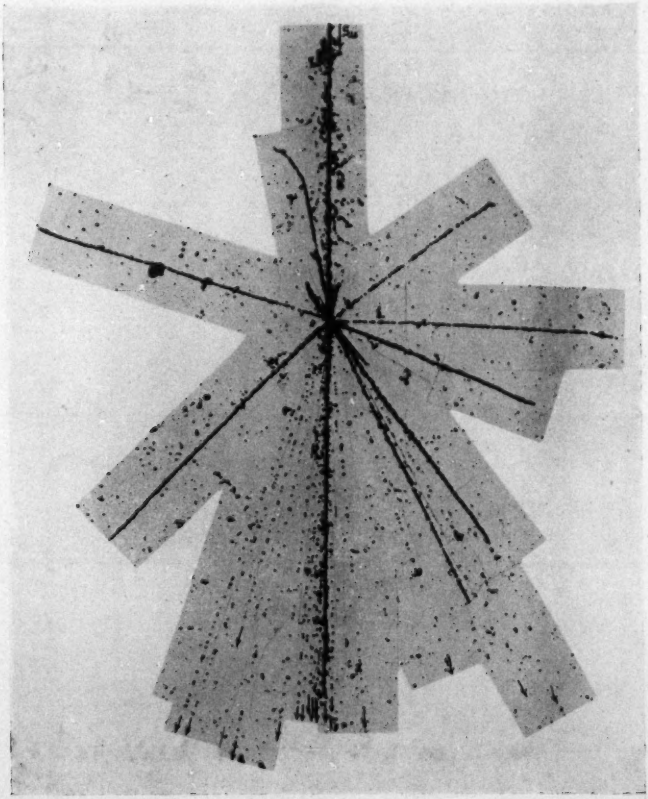


PLATE XI.

enters the emulsion, top left, and reaches the end of its range, bottom right. The break in the track
emulsions. The decrease in the range of the δ-rays as the velocity of the particle is reduced,
of electrons by the nucleus, are well displayed.



A sulphur nucleus ($Z=16\pm1$) collides with one of silver or bromide in the emulsion. As a result of the encounter a fluorine nucleus and twenty-five 'shower' particles—protons and π-mesons—emerge.

PLATE XII.

materials of intermediate atomic number. The experiments depended on the following considerations :—

If the disappearance of μ^--particles in materials of high atomic number is due to nuclear capture, there must be a rapid variation of the speed of the process with the value of Z, for while it is dominant for iron ($Z=26$), for carbon it is negligible in comparison with spontaneous decay. It was therefore anticipated that for materials with some intermediate value of the atomic number, the two competing processes would be of approximately equal importance. This would have the result that, first, the average life-time of the particles should be reduced in comparison with that observed in light elements; and second, that the proportion of mesons producing electrons would be reduced as the result of the removal of some of the particles by nuclear capture. On the other hand, if

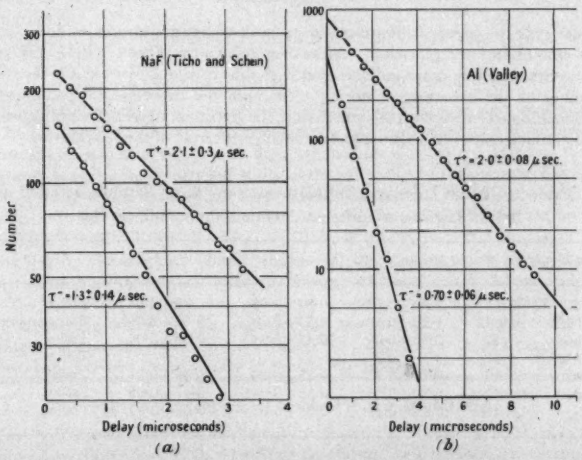


Figure 20. Decay of μ^--particles in (*a*) NaF and (*b*) carbon.

the effects were due to an acceleration of the decay, whilst the effect on the life-time would be indistinguishable from that expected on the alternative hypotheses, the proportion of mesons observed to produce decay electrons should be the same as that observed in light elements.

A typical result showing the variation of the number of electrons as a function of the time of delay for mesons stopped in (*a*) sodium fluoride and (*b*) aluminium is shown in Figure 20. It will be seen that the life-time in the heavy material is approximately half that of the positive particles, and that the number of electrons emitted per unit time, for zero time of delay, is also reduced.

Values of the life-time of μ^--particles stopped in various materials of different atomic number are summarized in Table 2. In addition to the values of the life-time τ^-, the Table gives the ratio, τ^-/τ^+, of this quantity to that for the positive

particles, and the fraction f of the negative mesons which decay with the emission of an electron. $1-f$ therefore represents that fraction of the negative μ-mesons which interact with nuclei.

These results establish the fact that the μ^--particles stopped in solid materials are in general removed by two competing processes, and the number of particles removed from an original sample in time δt can be written

$$\delta N = -(k_a + k_0)N\delta t,$$

N being the number of surviving particles, and k_a and k_0 'decay constants' corresponding to nuclear absorption and spontaneous decay, respectively. It follows that

$$f = \frac{k_0}{k_a + k_0} = \frac{\tau_a}{\tau_f},$$

where τ_a is the apparent life-time due to both competing processes and τ_f is the life-time of the free particles. It has been shown by Wheeler (1947, 1949) that theoretical considerations indicate that k_a should vary as Z^4, where Z is the atomic number of the absorbing material. Although the experimental results are in reasonable agreement with this prediction, the statistical weight of the observations is not sufficiently great to provide a detailed verification of the relationship.

Table 2. Mean Life-time (in micro-seconds) of μ^--Particles arrested in Solid Materials of different Atomic Numbers (*after* Rossi)

Substance	Z	Life-time (τ^-)	τ^+/τ^-	f	Author
O	8	$1{\cdot}89 \pm 0{\cdot}15$	$0{\cdot}87 \pm 0{\cdot}08$	$0{\cdot}83 \pm 0{\cdot}08$	Ticho
NaF	9, 11	$1{\cdot}23 \pm 0{\cdot}12$	$0{\cdot}57 \pm 0{\cdot}06$	$0{\cdot}60 \pm 0{\cdot}065$	Ticho and Schein
Mg	12	$0{\cdot}96 \pm 0{\cdot}06$	$0{\cdot}45 \pm 0{\cdot}04$	$0{\cdot}52 \pm 0{\cdot}04$	Ticho
Mg	12	$1{\cdot}1 \pm 0{\cdot}2$			Valley
Al	13	$0{\cdot}75 \pm 0{\cdot}07$	$0{\cdot}35 \pm 0{\cdot}04$	$0{\cdot}40 \pm 0{\cdot}04$	Ticho
Al	13	$0{\cdot}70 \pm 0{\cdot}06$	$0{\cdot}35 \pm 0{\cdot}035$	$0{\cdot}47 \pm 0{\cdot}05$	Valley
S	16	$0{\cdot}54 \pm 0{\cdot}12$	$0{\cdot}25 \pm 0{\cdot}03$	$0{\cdot}27 \pm 0{\cdot}03$	Ticho

It remains to consider the nature of the transmutations produced by the capture of μ^--mesons by the heavier nuclei. The evidence is now very strong that the process of capture is never, or very rarely, accompanied by the emission of charged particles. Several observers have recorded the tracks of μ-particles which are brought to rest in thin plates of heavy metals within Wilson chambers (Wang and Jones 1948, Chang 1949). In no case has a charged particle which could be identified unambiguously as a proton been observed. Further, the evidence obtained in experiments with photographic plates indicates that none, or very few, of the μ^--particles arrested in silver bromide produce stars (Camerini *et al.* 1948, Franzinetti 1950). It is therefore concluded that the disappearance of the rest-mass of the μ-meson results in the emission of neutral radiations.

Counter experiments by Piccioni (1948) provide convincing evidence that the nuclear absorption of the μ^--particle does not lead to the emission of photons. On

the other hand, it has been shown by Sard *et al.* (1948, 1949) that the process does lead to the production of neutrons. These authors find about one neutron for the nuclear absorption of every μ-meson. The simplest explanation of these results is that in the process of nuclear capture the meson interacts with one of the protons to produce a neutron and a neutrino in accordance with the equation

$$P + \mu^- \rightarrow N + \nu.$$

General theoretical considerations suggest that in such a process most of the energy will be carried off by the lighter of the two particles, so that the energy of recoil of the neutron will be small. It follows that even in those cases in which the neutron collides with other nucleons, the excitation energy of the nucleus is commonly insufficient to lead to the 'evaporation' of charged nucleons and the production of a detectable nuclear explosion (see Tiomno *et al.* 1949 and Chang 1949).

For a more detailed discussion of these considerations see Rossi (1948).

§6. ARTIFICIAL PRODUCTION OF MESONS

The first observations of the artificial production of mesons were made early in 1948 by Gardner and Lattes by means of the beam of fast α-particles generated by the 384 in. synchro-cyclotron at Berkeley, California. In the past two years, important progress in the detailed study of the properties of the particles has been made, a study which has recently been extended to include the generation of the π-mesons by fast protons and energetic γ-rays. In addition to providing the favourable conditions for the determination of the mass and life-time of the π-particles by the methods already discussed, these experiments give detailed information about the production of mesons by particles of relatively low energies, near the 'threshold' for π-meson production, where only a small fraction of the nuclear disintegrations are accompanied by the emission of mesons. It will therefore be convenient to consider the results obtained in these experiments before those derived from a study of the cosmic radiation.

6 (i). *Excitation Curve for the Production of Mesons*

The apparatus of the type shown schematically in Figure 5 (*a*) has been employed by Jones and White (1948) to determine the excitation-curve and the distribution in energy of the π^--particles produced by fast α-particles. The relative probability of producing π-particles, as a function of the energy of the primary α-particles, was determined by making exposures with the apparatus placed at different distances from the centre of the vacuum chamber of the cyclotron. In standard conditions of operation, the energy of the particles generated by the machine varies approximately as the square of the radius of the orbit, and by a suitable location of the target, particles of any desired energy can be chosen for experiment.

For each value of the α-particle energy, the number of mesons with kinetic energy in the interval from 2 to 10 MEV., emitted from the target in a direction within 45° of that of the primary particles, was determined by observations on the number of tracks per unit area of the plates recorded in an exposure of standard duration. The results are shown in Figure 21, curve (*b*). Results of similar

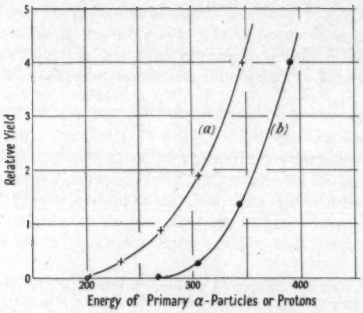


Figure 21. The variation in the yield of π^--mesons as a function of the energy of the generating particles : Curve (*a*) protons, curve (*b*) α-particles

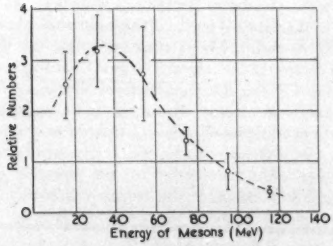


Figure 22. Distribution in energy of the mesons produced artificially by fast protons

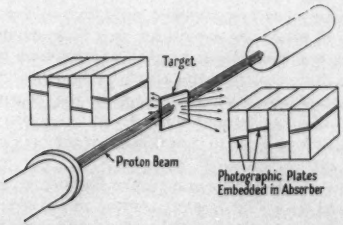


Figure 23. Apparatus of Richman and Wilcox for studying the generation of mesons by a beam of 345 MeV. protons emerging from the vacuum tank of the synchro-cyclotron

experiments on the generation of mesons by fast protons, obtained by the same authors are shown in curve (*a*).

The cross section, per carbon nucleus, for the production of mesons of energy from 2 to 5 MEV. by 390-MEV. α-particles has been measured by V. Peterson who has obtained an approximate value of $3{\cdot}0 \times 10^{-32}$ cm² per unit solid angle per MEV.

6 (ii). *Energy Spectrum of the Mesons*

The distribution in energy of the π^--particles generated by 390-MEV. α-particles has also been determined by Jones and White by means of apparatus similar in principle to that illustrated in Figure 5 (*a*), but in which the mesons in an extended range of values of $H\rho$ can be simultaneously recorded. The 'energy spectrum' of the mesons ejected in directions inclined at less than 30° to that of the α-particle beam can thus be determined.

Preliminary results have recently been obtained by Richman and Wilcox (1949) for the energy spectrum of the mesons produced by 345 MEV. protons. The apparatus employed is shown schematically in Figure 23. In these experiments the energy of the mesons has been deduced from the thickness of the absorber which they must have traversed in order to reach the particular photographic emulsion in which they were arrested. The observations shown in Figure 22 display a distribution in energy similar in form to that observed for the π-particles created by protons of great energy in the cosmic radiation : see § 7 (iii). The results have not been corrected to take account of the nuclear collisions of the π-particles in the absorbing material : see § 7 (v).

Cartwright *et al.* (1950) have recently described experiments on the positive and negative π-mesons emitted in the direction of the incident radiation as a result of the incidence of high energy protons on targets of graphite and of a suitable hydrocarbon. By employing targets containing the same numbers of carbon atoms, the effects due to hydrogen nuclei alone can be determined. It has thus been found that the distribution in energy of π-mesons—resulting from the interaction of 345 MEV. protons with protons, and emitted in the forward direction—shows a pronounced maximum at an energy of about 70 MEV. This value is near the maximum possible for the π-mesons as determined by the conservation laws. The authors direct attention to the possibility of extending their observations to include a detailed study of the distribution in energy of the mesons emitted in any direction from a target bombarded by protons or α-particles of known energy.

Reference has been made in § 5 (iv) to the measurement of the life-time of μ^+-particles, arising from artificially produced π^+-particles, in which advantage is taken of the short 'pulses' of protons which are delivered by the synchro-cyclotron. In the case of the γ-radiation generated by an electron synchrotron, the duration of the pulse cannot be made sufficiently short to allow a similar technique to be employed. In this case, however, it has been possible to employ crystal counters to detect the successive instants of occurrence of (*a*) the arrest of a π^+-particle, (*b*) the emission of the μ^+-particle, if this is delayed by more than 3×10^{-8} sec., and (*c*) the emission of the decay electron. Preliminary measurements by this powerful method by Kraushaar *et al.* (1950) give the value for the mean life-time of the π^+-particles $\tau_{\pi^+} = 1{\cdot}65 \pm 0{\cdot}33 \times 10^{-8}$ sec.; and by Steinberger and Bishop (1950) for the decay of the μ^+-mesons, $\tau_{\mu^+} = 2{\cdot}16 \times 10^{-6}$ sec.

6(iii). *Ratio of the Numbers of π^+- and π^--Particles*

The direct observations of the production of π-mesons in photographic plates exposed to cosmic radiation, see §7(i), have shown that in the case of emitted particles of low energy, there is a large excess of negative over positive particles. This effect may be readily explained in terms of the Coulomb field of the parent nucleus in which the particles are created. If the energy spectra for the two types at their points of creation are similar, then, especially for particles of low energy, the nuclear charge will exert an important influence on the distributions actually observed, the positive particles being accelerated and the negative retarded by the Coulomb forces. Such effects should depend on the nuclear charge, and they have been studied by Barkas who has bombarded targets of elements of different atomic number by α-particles of energy 390 MeV. The relative numbers of positive and negative particles, π^+ and π^-, produced by the bombardment of the different elements are shown in Figure 24.

For mesons of greater energy, $E>50$ MeV., the effects of the Coulomb field will be small. In the case of the bombardment of matter by protons, there should, however, be a disparity in the numbers of positive and negative π-particles created. Thus, suppose a nucleus to be composed of N neutrons and Z protons. An

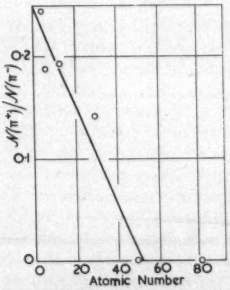


Figure 24. Ratio, $\mathcal{N}(\pi^+)/\mathcal{N}(\pi^-)$, of the number of positive to negative π-mesons, of energy in the interval from 2 to 5 MeV., generated by the impact of 390 MeV. α-particles with targets of different atomic number Z (Barkas 1949).

incident proton can react with a nuclear proton, either particle transforming to a neutron and leading to the creation of a positive meson: $\mathrm{P}+\mathrm{P}\rightarrow\mathrm{P}+\mathrm{N}+\pi^+$. On the other hand, if the primary proton interacts with a neutron, there are two possibilities for the production of charged mesons represented by the equations

$$\mathrm{P}+\mathrm{N}\rightarrow\mathrm{N}+\mathrm{N}+\pi^+,$$

$$\mathrm{P}+\mathrm{N}\rightarrow\mathrm{P}+\mathrm{P}+\pi^-.$$

If these two reactions are equally probable, the numbers of positive and negative mesons produced should be in the ratio $(2P+N)/N$. If the nuclei are of atomic weight A, charge Z, we may therefore write

$$\mathcal{N}(\pi^+)/\mathcal{N}(\pi^-)=(A+Z)/(A-Z). \qquad \ldots\ldots(10)$$

Experiments by Bradner and Jones (1950) indicate that the ratio of the numbers of positive to negative mesons of energy between 50 and 70 MEV., produced by the impact of 345-MEV. protons in carbon, is approximately equal to five, instead of three as indicated by the formula (23).

6(iv). *Experiments with γ-Radiation*

Similar experiments to those with protons, but employing γ-radiation produced by the impact on matter of the electrons produced by a 335-MEV. electron synchrotron, have been made by White, McMillan and Peterson (1950). The results obtained with the apparatus shown diagrammatically in Figure 25 indicate that the directions of emission of the π-particles are distributed isotropically The. distribution in energy of the mesons created in carbon appears to be similar to that produced by fast protons, for it extends up to 150 MEV. and has a maximum at 35 MEV. The cross section for the creation of a meson is found to be 5×10^{-28} cm^2

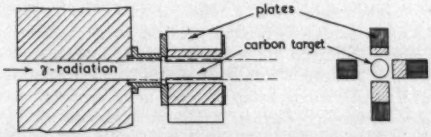


Figure 25. Apparatus for the study of the production of mesons by energetic γ-radiation (White, McMillan and Peterson 1950). The distributions in energy of the π^+- and π^--particles are deduced from the amount of absorbing material they are able to traverse before being brought to rest in photographic emulsions (Barkas 1949).

per steradian per carbon nucleus. In the case of production of mesons by photons, the negative particles are more numerous than the positive:

$$\mathcal{N}(\pi^-)/\mathcal{N}(\pi^+) = 1{\cdot}7 \pm 0{\cdot}2.$$

The above results appear to have an important bearing on the theory of the meson, for it has been shown by Brueckner (1949) that the experimental curves for the distribution in angle and energy of the emitted mesons can be best accounted for in terms of a pseudo-scalar theory. In contrast with the observations, scalar theory predicts a strong anisotropy in the angular distribution. Further, a comparison of the theoretical predictions with the observations appears to exclude the possibility that the mesons have spin 1.

6(v). *Production of Mesons by Energetic Neutrons*

Beams of neutrons of high energy are produced when fast protons are incident on matter, as a result of charge exchange with the nucleons of the struck nuclei. Neutrons of energy approximately 270 MEV. are thus created by the 340-MEV. proton beam from the synchro-cyclotron, and they may be allowed to emerge from the concrete shielding round the machine through suitable channels. By allowing the neutrons to pass through photographic plates, disintegrations can be recorded. About one in ten thousand of the resultant 'stars' shows the creation of mesons of short range, a phenomenon frequently observed in plates exposed to the cosmic radiation—see Smith *et al.* (1950).

§7. CREATION OF π-MESONS BY COSMIC RADIATION

7(i). *Ejection of π-Particles of Relatively Low Energy during Nuclear Explosions*

Shortly after the discovery of the π-mesons, nuclear explosions were observed in photographic plates which were accompanied by the emission of σ-mesons, particles which we have seen can now be identified as π^--particles (Lattes *et al.* 1947 b). A photomicrograph of an event of this type, of which many hundreds of examples have now been observed, is shown in Plate IV.

The analogous process in which a π^+-particle is emitted which comes to rest in the emulsion is much rarer, and only few examples have been reported (Leprince-Ringuet 1949, Powell 1949). In the example, of which photomicrographs are shown in Plate V, a π^+-particle is emitted during the explosive disintegration of a nucleus. It reaches the end of its range and emits a μ-meson which is also stopped in the emulsion, and which decays in turn to produce an electron.

The inequality between the observed numbers of 'ejected' π^+- and π^--particles of low velocity can be explained as a consequence of the Coulomb field of the nucleus in which the particle is created. In the early experiments, an ejected meson could only be identified if it reached the end of its range in the emulsion, so that any secondary effects which it produced could be distinguished. The probability that a particle will be brought to rest in the emulsion decreases rapidly with its velocity of emission. Further, any positive particle will be repelled from the parent nucleus as a result of the Coulomb forces, and will therefore necessarily have an energy of at least a few MeV. On the other hand, negative particles must escape against an attractive force and may therefore emerge with low velocity. Effects of this type are accentuated by the rapid increase in the range of a particle with increase in its initial velocity, and as a result the probability of observing an event of this type shown in Plate V is very small.

No observations have been reported which indicate that slow μ-particles can be created in nuclear interactions, and recent experiments by Piccioni (1950) and Fowler (1950), which are described in a later paragraph, make it almost certain that at least the majority of the mesons which are created in the form of 'showers' of fast penetrating particles are made up of π-particles. Both the mode of production of the π-mesons, and the features of the nuclear capture of the particles already discussed, are consistent with the view that they have a strong interaction with nucleons. In this respect, therefore, they are much more closely analogous to the Yukawa particles than the μ-mesons which, we have seen, have a very weak interaction with nucleons.

Although the π-particles appear to emerge from nuclei as the primary products of the interactions between nucleons, it can be objected that there is no evidence from the cosmic-ray experiments that they do not arise as the products of decay of a very short-lived 'primary' particle of even greater mass—that the observed chain of processes π-μ-e represents only the last stages of a longer succession of spontaneous transmutations. It is difficult to exclude this possibility in all cases because, if the life-time of such a postulated particle were less than 10^{-14} sec., the distance it would travel before decaying would be less than a few microns. Even with the photographic method—which, for this particular purpose, has a more extended time scale than any other—the recognition of the independent existence of the parent particle would, in such circumstances, appear to be impossible. Although a final opinion is not yet possible, the observations of the artificial

production of π-particles at Berkeley make it almost certain that π-mesons can be created directly and singly in nuclear collisions. The evidence is based on the following considerations :

The original observation of the artificial generation of π-mesons was made with a beam of α-particles of energy approximately 360 MEV. At a first glance it might be assumed that these particles would be essentially equivalent to four nucleons, each with an energy of about 90 MEV. In the collision of such a nucleon with a neutron or a proton, the principles of the conservation of momentum and mass-energy require that an energy of not more than 45 MEV. shall disappear as a result of the creation of particles of finite rest-mass. It might therefore be anticipated that it would be impossible for the 360 MEV. α-particles to create new particles with a rest-mass greater than about $90m_e$. Experiments showed, however, that π-particles were created by the bombardment of matter by α-particles of even lower energy, the 'threshold' for the process occurring at about 300 MEV. (see Figure 21).

This apparent contradiction arises because no account is taken in such speculations of the internal motions of the nucleons composing the α-particle and the nucleus with which it collides. If reasonable estimates of the magnitude of these internal velocities are made—and if it is assumed that meson production occurs in those most favourable cases in which the relative velocity of the interacting nucleons has the greatest possible value—it is found that the observed energy 'threshold' is in satisfactory agreement with that expected for the creation of a particle of mass $300m_e$ (McMillan and Teller 1947). It would, however, be difficult to explain the creation of particles of appreciably greater mass, particles which could be regarded as the parents of the π-particles. The results therefore strongly suggest that the π-particles can be created directly in nuclear interactions.

7 (ii). *Production of 'Showers' of Penetrating Particles*

For many years, following the pioneer work of Wataghin *et al.* (1940), Jánossy (1941) and other workers, it has been known that 'showers' of penetrating particles are sometimes created in the atmosphere by cosmic radiation. Characteristic apparatus for detecting events of this type is shown in Figure 26. It detects the simultaneous discharge of a number of independent Geiger counters, so arranged in massive lead blocks that the phenomenon cannot be due to cascade 'showers' of electrons and photons.

Observations with the expansion chamber, made by Fretter and others, appear to provide evidence of the mode of origin of these penetrating showers. Charged particles are observed which are able to penetrate several centimetres of lead. These particles sometimes interact with nuclei in one of the lead blocks, and from it groups of particles emerge which, since they are able to penetrate lead plates without producing cascade showers, cannot be electrons. It has commonly been assumed that these events correspond to the interaction of fast nucleons with nuclei, leading to the production of many mesons. The observed processes are reminiscent of those visualized by Heisenberg, Heitler and Jánossy, Oppenheimer and others in their theories of meson production.

Until recently, three important features of the phenomena have remained obscure. First, the nature of the shower particles has not been established, although it has been suggested that they are composed of a mixture of protons and mesons, the protons being projected with great energy in collisions with the primary particle

and the mesons being generated in nucleon–nucleon collisions during the penetration of the nucleus by the incident particle. Secondly, many mesons are often created in a nuclear interaction, and it has not been clear whether all, or many of them, commonly arise as a result of a single nucleon–nucleon collision, or, alternatively, whether they are usually produced singly in a succession of collisions of the primary particle with the nucleons lying near its line of motion in the traversal of a nucleus.

The first of these alternative theories of the origin of the 'showers' of penetrating particles is referred to as 'multiple production', and has been advocated by Heisenberg (1949) and Lewis *et al.* (1948). It is visualized that in a close collision between two nucleons a large fraction of the total energy available in the encounter may be emitted in the form of radiation as heavy 'quanta'—the limit to their

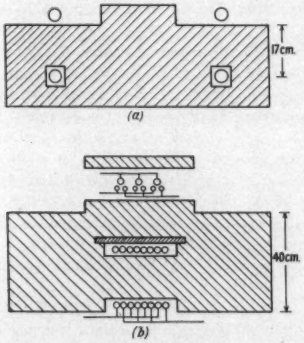


Figure 26. Apparatus of (*a*) Wataghin and (*b*) Jánossy for the observation of showers of penetrating particles.

number being imposed only by the requirements of the conservation laws of mass-energy and of momentum. On the other hand, Heitler and Jánossy (1949)—while not regarding the possibility as excluded that multiple production can sometimes occur—consider that the dominant process is that of 'plural' production—the creation, that is, of mesons singly in a succession of encounters of the primary nucleon in passing through the nucleus.

A third feature observed in connection with the penetrating showers is that they are frequently observed to be accompanied by soft radiation—by electrons and photons which 'multiply' in traversing the lead plates of the apparatus and produce characteristic showers of 'soft' radiation. Until recently it has not been clear whether this soft radiation is also created in the nuclear encounter, together with the penetrating particles, or whether it is of secondary origin.

The weight of the attack on these problems has recently been greatly increased by the observation of phenomena of the same type in 'electron-sensitive' photographic emulsions (Brown *et al.* 1949). Characteristic photo-micrographs of events of this type are reproduced in Plates VI, VII and VIII. It is an important

feature of the new observations that the 'shower' particles are observed at their points of creation, and that secondary processes associated with their production, such as the subsequent 'evaporation' of the nucleus, the ejection of protons and α-particles of relatively low energy, can also be observed.

7 (iii). *Nature of the Shower Particles*

The first definite evidence on the nature of the shower particles was obtained in experiments of great ingenuity carried out by Piccioni (1950), using apparatus shown diagrammatically in Figure 27. There are three trays of counters separated by lead blocks, A, B and C, and the third tray can be surrounded by additional blocks, D and E, which can be either of carbon or of sulphur. The method takes

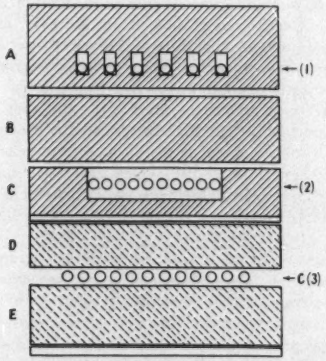


Figure 27. Apparatus of Piccioni for the identification of 'shower' particles.

advantage of the difference in the behaviour of μ^- and π^--particles arrested in light and heavy elements. Thus while π^--particles interact with nuclei when stopped in either carbon or sulphur, the μ^--particles are absorbed by nuclei only in the heavier element; in carbon they decay with the emission of an electron (§ 5 (v)).

Piccioni showed that when single penetrating particles traverse the apparatus and are stopped in D or E, delayed coincidences due to the discharge of a counter in tray (3) can be observed. The ratio of the number of such events observed per unit time in the case of the sulphur and carbon absorbers, $N_C/N_S 1{\cdot}8$. This result is in good accord with that to be expected for μ-particles; the electrons are produced by the arrest of both positive and negative particles in the case of carbon, and of positive particles only in the case of sulphur. The result is a little different from 2 because of a well-known excess of positive particles among the μ-mesons of the penetrating component.

In contrast with the effects due to isolated penetrating particles, nuclear explosions occurring in the lead block A which are accompanied by the emission of shower particles can result in the discharge of two or more of the counters in tray A. It was found further that coincidences of two such counters with one or

more in B are almost all due to such locally produced showers. In the case of these events, the delayed coincidences in the absorber C were found to occur with almost equal frequency whether D and E were of graphite or of sulphur. This is consistent with the assumption that the locally produced mesons are all π-particles, for in this case, both in light and heavy elements, only the positive particles will transform to give μ-particles and then decay electrons; and the results are inconsistent with the assumption that more than a small fraction of the locally produced mesons are μ-particles.

Whilst Piccioni's experiment provided very strong evidence for the production of π-mesons in nuclear explosions, it gave no indication of their frequency of occurrence among the shower particles, or of their distribution in energy. Most of the particles of the 'penetrating showers' are ejected with velocities in the relativistic region, and it is well known that this renders difficult the determination of their rest-mass. Certain conclusions can, however, be drawn from the simultaneous observation of the curvature of the trajectory of a particle in a Wilson chamber operated in a magnetic field—which gives a measure of its momentum—and of its specific ionization as determined by 'drop-counts'. Such experiments are difficult to make with precision—see § 3 (v)—and they have recently been supplemented by experiments with photographic plates, determining the momentum of the particles by observations on the deviations in the tracks due to multiple Coulomb scattering (Camerini *et al.* 1949, Occhialini 1949, Goldschmidt-Clermont 1950). The scope of the method has been greatly extended by the development of a very simple and rapid method of measuring the scattering (Fowler 1950).

A typical result obtained by Fowler is shown in Figure 10, in which the grain-density for each track is plotted against the scattering parameter, $\bar{\alpha}$, the average deviation per unit length. The tracks on which measurements were made were chosen only because their lengths were greater than 3 mm., and irrespective of their relation to the 'stars' with which they were associated. Some of the tracks were therefore due to primary particles which produced nuclear explosions, but these occur rarely compared with those of secondary particles.

The results shown in Figure 10 show that a large proportion of the particles which produce tracks with a grain-density near the minimum value are less massive than the proton. Further, the average value of the mass of those of the mesons with a specific ionization greater than the minimum value—as deduced from the displacement of the points from the corresponding curve due to protons—is $283 \pm 7\, m_e$ (see Figure 11). They may therefore be identified as π-particles. The observations prove that if any μ-mesons or electrons are emitted from the nuclear explosions with an energy less than 150 MEV., they are fewer than 2% of the π-mesons with energies in the same range; and the observations strongly suggest that, apart from ejected nucleons, π-particles form the great majority of the charged particles directly produced in nuclear reactions.

It has been known for many years (Blackett and Wilson 1938) that most of the electrons of the soft component at sea level have energies less than 200 MEV. The results of the new experiments therefore suggest that most of the electrons do not arise as a direct result of nuclear interactions, but are of secondary origin. From these experiments alone the possibility is not excluded that γ-radiation is created in the nuclear interactions, and this could also lead to the production of cascade showers. We shall see in a later paragraph, however, that the soft radiation

accompanying the penetrating showers can be accounted for in other ways—by assuming that it arises from γ-radiation formed by the decay of neutral mesons which are created together with the showers of charged π-mesons in nuclear interactions.

7 (iv). *Distribution in Energy of the Mesons of the Penetrating Showers*

The methods described in the preceding paragraph can be employed to determine the distribution in energy of the mesons created in the disintegrations produced by cosmic-ray particles of great energy. Results thus obtained by Camerini *et al.* are shown in Figure 28. These results may be compared with the distribution in energy of the mesons of the penetrating component of the cosmic radiation. From observations of the intensity of slow μ-mesons at

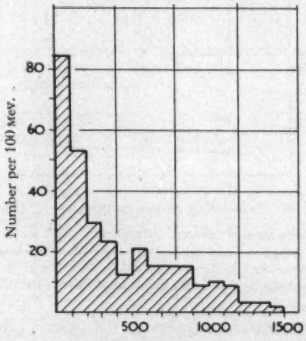


Figure 28. Distribution in energy of the mesons among the shower particles observed at an altitude of 70,000 ft.

different depths in the atmosphere, and the energy distribution of the particles at sea level, Sands (1950) has computed the distribution in energy of the particles at their points of production. If it is assumed that the μ-mesons arise by the decay in flight of π-mesons, Sands' results can be transformed to correspond to the energy distribution of the parent π-particles at their points of creation, The corresponding curve is found to be in good agreement with the distribution shown in Figure 28. This result gives strong support for the view that the μ-mesons of the penetrating component arise by the decay in flight of the π-mesons forming in nuclear explosions occurring in the atmosphere.

In favourable cases the energy of the primary particles producing nuclear explosions in photographic emulsions can also be measured. By a statistical study of a large number of events it is then possible to determine the average energy of the primary particles which produce disintegrations in which different numbers n_s of fast ' shower ' particles are created. Further, a comparison can be made between the average energy $\bar{E}_p$ of the primary particles producing showers of a given multiplicity n_s, the total energy represented by the emitted mesons, $n_s\bar{E}_s$, and that of the ejected nucleons and α-particles forming the ' star '.

The results of a preliminary study of this type by Camerini *et al.* are represented in Figure 29. The results indicate (*a*) that the average energy of the mesons changes very little with the multiplicity n_s, the greater the energy of the ' primary ' particle, the greater the number of mesons of approximately the same kinetic energy that are produced, and (*b*) that a fraction of the energy of the primary particles is unaccounted for. It is believed that this missing energy contributes to the creation of neutral mesons which commonly escape observation (see § 8).

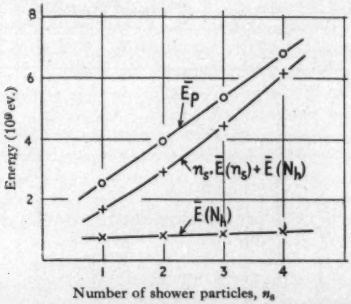


Figure 29. Energy balance in stars of different multiplicity : $\bar{E}_p$ is the mean energy of the primary protons, n_s, $\bar{E}_s$ the average value of the total energy(mc^2) of the shower particles and $\bar{E}(N_h)$ the mean energy represented by the star. It will be seen that some energy is missing, and it is suggested that this may be due to the creation of neutral mesons.

7 (v). *Nuclear Interactions of the Shower Particles*

A problem which has had an important bearing on the nature of the shower particles is the frequency with which they make nuclear collisions in passing through matter. The μ-mesons are able to penetrate to great depths underground because they interact with nuclei very rarely. If, therefore, the argument runs, many of the particles emitted in ' showers ' are μ-mesons, the average length of path before making a nuclear collision should be much greater than the value calculated on the assumption that the particles interact with every nucleus through which they pass, the cross section for collision then being given, approximately, by the geometrical values: $\sigma = A^{2/3} \times 6{\cdot}8 \times 10^{-26}\,\mathrm{cm}^2$. On the other hand, if the shower particles, near their points of production, are π-particles, the average path-length for nuclear collision should not be widely different from the value deduced from a ' geometrical ' cross section.

About 80% of the shower particles are now known to be π-mesons. The significance of the observations of the average path length has therefore changed, and it now bears on the important question of the strength of the interaction of π-mesons with nucleons. Examples of the secondary nuclear interactions of ' shower ' particles, as observed in photographic emulsions, are shown in the photomicrographs reproduced in Plates VI and VII.

Determinations of the average path length of the shower particles have been made in a number of experiments with widely differing results. The method most commonly adopted has been to observe the creation of the shower particles

as a result of nuclear explosions occurring in lead plates within Wilson chambers. The shower particles subsequently traverse other lead plates within the chamber, and estimates are made of the fraction of the particles which fail to emerge from these subsequent absorbers as a result of nuclear interactions. The results obtained by a number of experimenters with this method are included in Table 3. In order to be able to compare the results of different experiments employing different absorbers, the Table includes the ratio of the observed 'mean free path' to the value calculated on the assumption that the nuclei of the absorber present a cross section for interaction equal to the geometrical value.

Table 3

Authors	Method	Result	Observed M.F.P. / Geom. value	Energy range (mesons)
Piccioni (1950)	Counters at mountain altitudes	1200 gm.cm^{-2} Fe	14·0	>400 Mev.
Fretter (1949)	Cloud chamber with Pb plates at mountain altitudes	750 gm.cm^{-2} Pb Uncorrected	4·7	>150 Mev. $n_s \geqslant 2$
Lovati, Mura, Salvini and Tagliaferri (1949)	Cloud chamber with Pb plates at mountain altitudes	300±100 gm.cm^{-2} Pb Corrected	1·9	>150 Mev. $n_s \geqslant 2$
Brown and McKay (1950)	Cloud chamber with Pb plates at mountain altitudes	316±70 gm.cm^{-2} Pb Corrected	2·0	>150 Mev.
Butler, Rosser and Barker (1950)	Cloud chamber with Pb plates at sea level	400 gm.cm^{-2} Pb Uncorrected 200 gm.cm^{-2} Pb Corrected	2·5 2·5 1·4	Average energy of shower ~7000 Mev.
Harding and Perkins (1949)	Photographic plates exposed under ice	120 gm.cm^{-2} Ice	2·0	>100 Mev.
Camerini, Fowler, Lock and Muirhead (1950)	Electron-sensitive plates exposed at high altitudes	100 gm.cm^{-2} emulsion	1·1	>150 Mev.

Observations with photographic plates which have led to values for the mean free path substantially smaller than that obtained with Wilson chambers have been described recently by Camerini *et al.* (1950). These experiments indicate that the average path length, both for fast protons and for π-mesons, is close to the 'geometrical' value. These authors show that the π-mesons commonly produce small 'stars', and they suggest that the experiments with Wilson chambers are complicated by the fact that the particles emitted as a result of the secondary interactions occurring within the lead plates will often not emerge from the lead. In such cases the secondary interactions will escape observation, so that an over-estimate of the average path-length between collisions will result.

Whilst final values for the cross section for the collision of π-mesons with nuclei are not yet available, the evidence suggests that they have a high probability of interacting with every nucleus which they penetrate. This question is of great importance for a detailed study of the processes leading to the production of the 'showers'. Especially in heavy nuclei, secondary interactions within the parent nucleus of mesons produced in nucleon–nucleon collisions may play an important role.

7 (vi). *Multiplicity of Production of Mesons*

Reference has already been made to the problem of the multiplicity of production of mesons in nucleon–nucleon collisions. The question is of decisive importance, and in the discussions on the subject which took place at the International Conference at Como in September 1949 it was generally agreed that a decisive experiment would consist in observing the generation of many mesons in the interaction of a proton or neutron of great energy with a hydrogen nucleus.

Suppose a fast proton interacts with second proton at rest. In the absence of the production of charged mesons, the collision will lead to the projection of the struck proton, so that the collision will give rise to a forked track. If, alternatively, production of charged mesons occurs, the particle must be produced in pairs in order to provide for the conservation of charge, so that the number of secondary particles will be two, four, six, etc. This conclusion is not altered if charge exchange occurs so that one or both of the original protons involved in the collision emerge as neutrons. The total charge carried by the particles involved in the collision, including that of the original fast particle, must therefore be odd. The same result is obtained if a fast neutron is assumed to collide with a proton, for in this case the total number of secondary particles is odd and the primary particle produces no track.

Events of the type discussed in the previous paragraph if they occur in photographic emulsions could be expected to produce 'stars' in which all, or nearly all the tracks have a grain-density equal to the minimum value for fast particles of charge $|e|$. They would therefore commonly escape observation except in the examination of the plates under high magnification. Following the Como discussions, several observers directed their attention to making such a search, and already a number of examples have been found which appear to correspond to the expected process.

In a private communication, Professor Heisenberg informs me that he and his colleagues have found an event in which a fast particle produces a star from which the tracks of six particles diverge. Five of these tracks are of minimum grain-density, and the sixth is that of a slow π^--particle. These features are consistent with the assumption that the event corresponds to a proton–proton collision in which four π-mesons are created. A similar event is shown in Plate IV, in which, in addition to the π-particle which produces a disintegration, five fast particles and a single heavily ionizing particle are produced. In another event observed in Bristol, the star is made up of a total of eight tracks, one primary and seven secondary all of which are of minimum ionization. In this case, the event cannot be finally identified because the total number of tracks is even, but it is reasonable to assume that a ninth track is associated with the star which cannot be distinguished because of an unfavourable direction of motion of the particle. It is known that steeply dipping tracks are sometimes difficult to detect.

Further observations, which appear to be very difficult to explain in terms of pure plural production, have been made on the interaction with nuclei of α-particles of great energy. Thus Bradt, Kaplan and Peters (1950) have observed such a collision in photographic plates exposed at altitudes of the order of 90,000 feet by means of free balloons. The α-particle produces a nuclear explosion which is accompanied by the creation of a narrow 'jet' of about 23 fast particles most of which we may assume to be π-mesons. A similar event in which an α-particle produces 35 fast particles is shown in Plate VIII. It will be seen that in this case only three heavily ionizing particles are emitted. The evidence given in a later section strongly suggests that the charged mesons produced in these events are accompanied by neutral meons in considerable numbers. It appears to be very improbable that a process of pure plural production could lead to the production of about 40 mesons in a nuclear collision, and that only three protons should emerge from the encounter. Further, it seems probable, to judge from the preliminary evidence which has been accumulated up to the present, that the α-particles of great energy have an appreciable probability of producing effects of the type shown in Plate VIII, and that the ejection of only a relatively few charged nucleons is not usual. It appears therefore that the event is not exceptionally rare and that it does not display unusual features.

The evidence at present at our disposal appears to indicate that multiple production of π-mesons in individual nucleon–nucleon collision does occur; but the relative importance of single and multiple production in the creation of shower particles and the detailed understanding of the process occurring within the nucleus remain to be elucidated. For this purpose the determination of the mean path length of π-particles between nuclear collisions is insufficient. It allows only an upper limit to be given for the range of the forces between a π-particle and a single nucleon, and it is the magnitude of this quantity which is decisive.

7 (vii). *Production of 'Showers' at great Depths Underground*

Recent experiments by George have shown that nuclear explosions accompanied by the emission of 'showers' of particles moving at relativistic velocities occur at considerable depths underground. The 'electron-sensitive' emulsions were prepared 30 m. underground, exposed for some time at the same depth, and processed. In this way the possibility that the phenomena are due to pre-exposures at sea level is completely excluded.

The plates prepared in this way record about 10^{-2} nuclear disintegrations per cm^3 per diem, and of these about one-fifth are accompanied by the emission of showers. In about a quarter of the events the track of a fast particle of charge e which produced the disintegration can be distinguished. The nature of the shower particles remains to be established, but the events appear to be closely similar to those observed at high altitudes, and it is reasonable to suppose, tentatively, that the shower particles are π-mesons.

It is very difficult to account for those disintegrations in terms of their production by energetic protons or π-particles which have penetrated underground without being absorbed as a result of nuclear interactions. On the other hand, at these depths, there is a flux of μ-mesons of considerable intensity. Further, recent experiments at Berkeley have shown that electromagnetic radiation can interact with nuclei to produce mesons. It is therefore reasonable to suggest that the

'showers' are due to the electromagnetic interaction of fast μ-mesons with nuclei. From the known flux of these particles and the rate of production of showers, the cross section for the process is found to be of the order of 10^{-29} cm^2. Many of the stars observed at the same depth, but unaccompanied by 'showers', may be attributed to π-mesons and fast nucleons arising as secondary particles from the 'showers' (George and Evans 1950).

§ 8. NEUTRAL MESONS

8 (i). *Transmutation of Neutral Particles into Quanta*

In the first section of the present report it was pointed out that if the basic features of Yukawa's theory are correct, the approximate equality of the forces between like and unlike nucleons appears to require the assumption of the existence of neutral as well as of charged mesons. Recent observations by Bjorkland *et al.* (1950) strongly suggest that such neutral particles, of rest-mass approximately

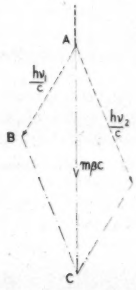


Figure 30. Momentum balance in the decay of a neutral meson into two photons.

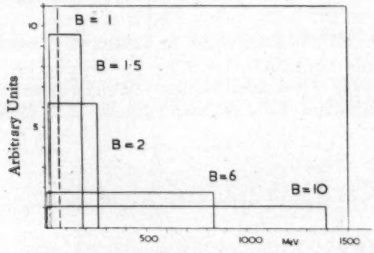
$300m_e$, are created in energetic nuclear collisions, that they are of short life-time, less than 10^{-11} sec., and transform into two γ-rays. Before considering the details of the experimental results, it will be convenient to establish the principal characteristics of the radiation to be expected from such a mode of decay of a material particle.

If a neutral meson, moving with a given velocity, suffers decay into two photons, the energy of the quanta will depend on their directions of emission relative to the line of motion of the parent particle. The problem is precisely analogous to the emission of radiation by a moving source; the wavelengths and intensity distributions are governed by the well-known Doppler principles. Let m_0 be the rest-mass of the particle, and $\beta\mathbf{c}$ its velocity; its momentum is then $m\beta\mathbf{c}$, where $m=m_0/\sqrt{(1-\beta^2)}$. Further, suppose that $\boldsymbol{h\nu_1}$ and $\boldsymbol{h\nu_2}$ are the energies of the two quanta, and that they are emitted in directions making angles θ and ϕ with respect to the line of motion of the neutral meson. The vector sum of the momenta of the two quanta, $\boldsymbol{h\nu_1/c}$ and $\boldsymbol{h\nu_2/c}$ (see Figure 30), must equal the momentum of the parent particle $m\beta\mathbf{c}$; $\mathbf{AB}+\mathbf{BC}=\mathbf{AC}$. By the conservation of energy, however, $\boldsymbol{h\nu_1}+\boldsymbol{h\nu_2}=m_0\mathbf{c}^2/\sqrt{(1-\beta^2)}=Bm_0\mathbf{c}^2=$constant. It follows

that $AB + BC$ is a constant, and the locus of B is an ellipse. It is easily shown that the energy of the quantum emitted at an angle θ is given by the relation

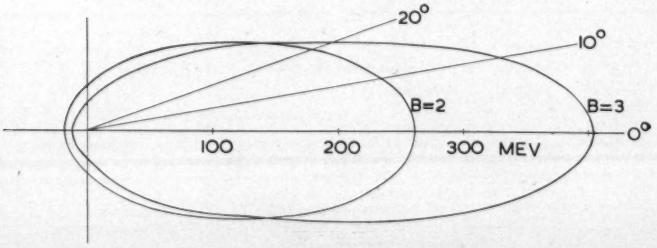
$$h\nu = \frac{m_0c^2}{2B(1-\beta\cos\theta)}.$$

Curves showing the calculated distributions of the photon energy for different directions of emission, and for different values of the total energy of the neutral mesons, Bm_0c^2, are shown in Figure 31 (*b*), the rest-energy of the meson, m_0c^2, being taken as 140 MEV.



(*a*)

Energy distribution of photons from a monokinetic beam of neutral mesons.



(*b*)

Characteristics of the radiation produced by the decay of neutral mesons into photons. (*a*) Spectral distribution of the radiation; (*b*) quantum energy as a function of the angle of emission. Curves for different values of B, where Bm_0c^2 is the total energy of a particle of rest-mass m_0.

Figure 31.

The maximum and minimum values of the energy of the γ-rays corresponding to emission in the forward and backward directions with respect to the line of motion of the parent meson are given by the relations

$$h\nu\,(\max) = B\frac{m_0c^2}{2}\left(\frac{1+\beta}{1-\beta}\right); \qquad h\nu\,(\min) = B\frac{m_0c^2}{2}\left(\frac{1-\beta}{1+\beta}\right).$$

Even for small values of the kinetic energy there is a considerable width in the spectral distribution of the emitted radiation. The calculated form of the spectrum, for different values of B, is shown in Figure 31 (*a*), each curve

representing the results for a homogeneous group of neutral mesons of given kinetic energy.

Finally, it can be shown that the intensity of the radiation emitted at an angle θ, and measured in terms of the number of quanta per solid unit angle, is given by the relation

$$I(\theta) = k/B^2(1-\beta\cos\theta)^2, \quad \text{where } k \text{ is a constant.}$$

The curves showing the distribution of intensity become strongly asymmetric for values of B greater than 2, with a marked tendency for the radiation to be emitted in directions inclined at only a small angle to that of the parent particle.

8 (ii). *Experimental Evidence for the Existence of Neutral Mesons*

In the experiments of Bjorkland *et al.* fast protons produced by the synchro-cyclotron were allowed to impinge on matter. When the generator is in operation, there is an intense background of radiation due to a variety of nuclear processes,

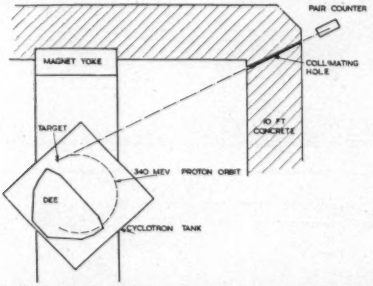


Figure 32. Disposition of apparatus for observations on the γ-radiation produced by the impact of 340 MeV. protons in matter.

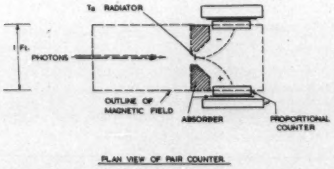


Figure 33. Apparatus for the determination of the intensity and quantum energy of γ-radiation. (*After* Bjorkland *et al.* 1950.)

and it is therefore surrounded by thick walls of absorbing materials. These are perforated with suitable channels so arranged that any radiation originating in the immediate vicinity of the target can escape to the outside (Figure 32). The emerging γ-rays were allowed to fall on the apparatus, shown in Figure 33, which consists of four sets of counters connected to record coincidences, and operated in a magnetic field.

If a photon of high energy falls on the thin lead radiator (see Figure 33) it has an appreciable probability of producing a pair of positive and negative electrons. These are deflected in opposite directions by the magnetic field, and may then pass through each of the two pairs of counters. A pair of electrons of opposite sign, and with energies in a certain range of values, can therefore lead to the simultaneous discharge of the four counters—can lead, that is, to a 'quadruple coincidence'. These coincidences give a measure of the intensity of the high energy γ-radiation emerging through the channel.

Estimates of the energy of the pairs of electrons can also be obtained by interposing metal sheets of suitable stopping-power between the counters of each pair. The greater the energy of the electrons, the greater the thickness of the absorbing sheets required to produce a given diminution in the rate of occurrence of the coincidences. Observations of the reduction in intensity produced by different absorbers therefore give a measure of the quantum energy of the γ-radiation.

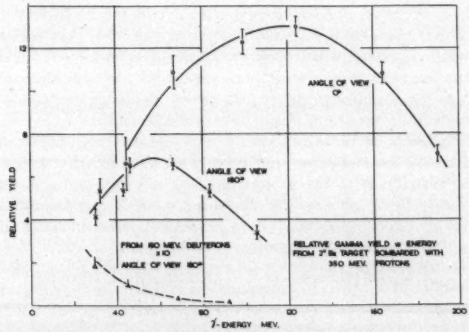


Figure 34. Characteristics of the radiation produced by the bombardment of carbon by protons of energy 340 MeV. The spectral distribution is shown for two directions of emission with respect to the line of motion of the generating particles.

Provision was made to observe the γ-radiation, arising as a result of the bombardment of the target by protons of great energy, in either of two directions. In one case the direction of emission of the radiation was nearly parallel to that of the primary protons, and, in the other, the radiation emerging nearly backwards was examined.

The most striking of the results obtained may be summarized as follows: (*a*) If the intensity of the γ-radiation emitted in a given direction is studied as a function of the energy of the protons bombarding the target, it is found that the production of radiation begins rather suddenly at a certain 'threshold' energy, and this 'threshold' occurs at nearly the same value as that for the production of charged π-mesons. The intensity of the emitted radiation increases rapidly with the increase of the proton energy above the 'threshold'. (*b*) The energy of the γ-radiation observed in a given direction has a certain 'spectral distribution', with a well-marked maximum (Figure 34). The average quantum energy is

considerably greater for the photons emitted forward in the direction of the proton beam than for those emitted backward. Further, the intensity of the radiation is greater in the forward direction.

These and other secondary features of the observations can be very simply explained if the γ-radiation is assumed to arise as a result of the creation of neutral mesons, together with charged mesons, in energetic nucleon–nucleon collisions. The neutral meson is assumed to be very unstable, with a life-time less than 10^{-11} sec., before decaying into two γ-rays. The assumption of a short life-time is necessary because it can be shown that the observed γ-radiation originates from points within 3 mm. of the 'target', and it follows that any parent neutral meson has a large probability of suffering spontaneous decay in traversing this distance.

The observed quantum energies and intensities observed in the experiments are in good accord with the view that the neutral mesons are emitted isotropically in the centre-of-mass system of pairs of colliding nucleons, one of which is provided by the incident particle and one by a nucleon in a target nucleus.

The tentative conclusions drawn by the authors from these experiments appear to have received decisive support from two recent results obtained at Berkeley. It has been shown by Steinberger *et al.* (1950) that the high energy photons from a target bombarded by energetic protons are produced in pairs. The quantum energies of these photons vary with the direction of emission with respect to the beam of primary particles in a manner consistent with their suggested mode of origin by the decay of neutral mesons.

Secondly, experiments on the capture of π^--mesons in high pressure hydrogen have been made by Panofsky *et al.* (1950). It has been found that the process of capture commonly leads to the production of a neutron and a neutral meson, the latter particle being detected by the observation of the pairs of γ-rays into which it decays. More rarely the meson produces a neutron and a single γ-ray. The two reactions can therefore be written

$$(a) \quad \pi^{-1} + \mathrm{H}^1 \rightarrow \mathrm{n}^0 + \pi^0; \qquad \pi^0 \rightarrow 2\hbar\nu (\hbar\nu \sim 70 \text{ Mev.})$$

and

$$(b) \quad \pi^{-1} + \mathrm{H}^1 \rightarrow \mathrm{n}^0 + \hbar\nu; \qquad (\hbar\nu \sim 140 \text{ Mev.}),$$

where π^0 represents a neutral meson.

It is clear from the known masses of three of the particles in equation (*a*) that the mass of the π^0-particle is less than that of the charged meson. Further, the experiments allow an estimate to be made of the inhomogeneity in energy of the γ-rays from reaction (*a*). We have seen that the extent of such an inhomogeneity will be very sensitive to the kinetic energy of the recoil of the neutral meson, and the observations show that the velocity of recoil is small. It follows that the mass of the π^0-particle is only a little less than that of the charged meson, and the preliminary results indicate that $m_{\pi^+} - m_{\pi^0} \sim 4\, m_e$, where m_e is the mass of the electron. It follows that $m_{\pi^0} \sim 270\, m_e$.

8 (iii) *Evidence for the Production of Neutral Mesons by Cosmic Radiation*

Reference has been made in § 7 (vi) to the observation by Bradt, Kaplan and Peters (1950) of a highly collimated 'jet' of shower particles produced in the collision of an α-particle of great energy with a nucleus. This event occurred in the emulsion of one of an assembly of many plates, and the 'jet' could be followed in its passage through a succession of emulsions.

Bradt *et al.* found that at a distance of a few centimetres from the nuclear explosions pairs of electrons were created among the charged particles of the shower. The electrons were of great energy and moved in directions inclined at only small angles to the 'axis' of the shower. We have referred to the pairs of particles as electrons. Although their nature remains to be established, it is reasonable to assume that they are indeed pairs of electrons created by γ-radiation. The authors observed nine such pairs within a distance equal to about a radiation-length from the nuclear explosion, and they concluded that the 56 shower particles must have been accompanied by about 35 energetic γ-rays.

We have seen that the recent experiments of Fowler (1950) prove that at least 80% of the charged particles in the penetrating showers are π-mesons, and that very few, and possibly none, are electrons. Further, serious theoretical difficulties are met if it is assumed that large numbers of γ-rays are produced directly in energetic nuclear collisions. On the other hand, the observations of Bradt *et al.* are readily explained if the γ-radiation which creates the pairs of electrons observed in their experiments is attributed to the spontaneous decay of neutral mesons produced in numbers comparable with that of the charged π-particles. A detailed analysis of the characteristics of the event has led to the conclusion that the mean life-time of the particles is less than 3×10^{-13} sec.

8(iv). *Mass and Energy of Neutral Mesons in Penetrating Showers*

Additional evidence bearing on the mode of origin of the γ-radiation present in the cosmic radiation, and a method which, in principle, makes it possible to determine the life-time of neutral mesons and to establish the independent existence of the particles, has been described recently by Carlson, Hooper and King (1950). These authors have determined the spectral distribution of the γ-radiation in the atmosphere at 70,000 ft. by measuring the energy of the individual pairs of electrons created by its passage through the emulsion. Assuming that this radiation also arises by the decay of neutral mesons created in nuclear interactions, and that the 'spectrum' has not been seriously modified by the presence of photons produced by 'bremsstrahlung', it is possible to deduce (*a*) the mass of the postulated neutral mesons, and (*b*) the energy 'spectrum' of these particles at their points of creation.

The method is based on the following considerations. Figure 31(*a*) shows the calculated distributions in energy of the photons produced by the decay of homogeneous groups of neutral mesons for different values of the total energy Bm_0c^2. For an inhomogeneous beam, in which the neutral mesons are distributed in an energy spectrum, there is a maximum in the distribution at an energy $\frac{1}{2}m_0c^2$. Further, suppose that the two values of the energy for which the intensity of the radiation is equal to any arbitrarily chosen value are E_1 and E_2. It can then be shown that $(E_1E_2)^{1/2} = m_0c^2/2$, whatever the chosen value of the intensity.

The above result shows that if the γ-radiation is produced by the decay of neutral mesons, the form of the 'spectrum' of the photons must display features consistent with the relation established above. The observed energy distribution at 70,000 ft., as determined by Carlson *et al.*, is shown in Figure 35. Five pairs of values of the energy intercepts, for different values of the intensity, have been determined, and the corresponding values of the quantity $(E_1E_2)^{1/2}$ are shown by circles in the figure. The degree of consistency of the values is remarkable, and the mean corresponds to a mass of $295 \pm 20\, m_e$ for the neutral meson.

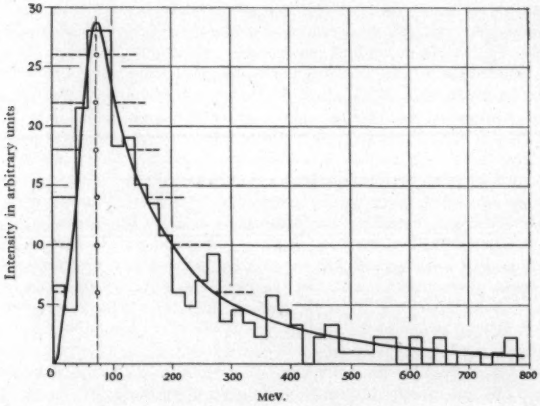


Figure 35. Spectrum of the γ-radiation in the atmosphere at an altitude of 70,000 ft. (Carlson, Hooper and King 1950).

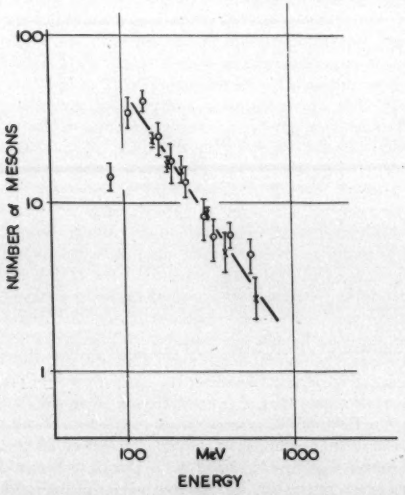


Figure 36. Energy distribution of the charged (○) and neutral (×) π-mesons in the atmosphere at 70,000 ft. The intensity scales are arbitrary (Carlson *et al.* 1950).

Using the value for the mass of the neutral meson thus determined, the observed γ-ray spectrum can be employed to deduce the distribution in energy of the neutral mesons at their points of creation. The result thus obtained is shown in Figure 36. Included in the same figure is the corresponding curve for the charged mesons observed to be created in 'showers' occurring in the same plates. It will be seen that there is no significant difference between the forms of the two distributions, an indication that the neutral and charged mesons are created in similar types of processes.

These observations provide convincing evidence that in the explosive disintegrations produced by protons and α-particles of great energy, the charged π-particles of 'showers' are accompanied by neutral mesons of the same type as those observed to be produced artificially. For the successful application of the method of Carlson, Hooper and King it appears to be essential to make observations at high altitudes where the γ-radiation has not been modified appreciably by the development of cascade showers.

8(v). *Life-time and Frequency of Production of Neutral Mesons*

The evidence for the existence of neutral mesons presented in the previous paragraphs, whilst very strong, is circumstantial. It is therefore a matter of importance to obtain, if possible, decisive evidence of the independent existence of the particles. For this purpose, and in order to measure the life-time of the particles, Carlson *et al.* have made observations on the pairs of electrons which are sometimes created, close to a 'star' in the emulsion, by γ-rays resulting from the nuclear explosion.

Suppose that a homogeneous beam of neutral mesons, each of rest-mass m_0, is moving with such a velocity that the total energy per particle is Bm_0c^2. We have seen that for values of $B>2$ there is a pronounced tendency for the emitted γ-radiation to preserve the direction of motion of the parent particle. As a consequence of these features of the postulated mode of decay of the neutral mesons if follows that the line of motion of the photons will, in general, be inclined to that of the parent particle. Further, in the case of photons in the range of energies with which we are concerned, the two electrons into which each eventually converts will form a very narrow 'pair' with a divergence of the order of 0·1°. They will thus give an indication of the line of motion of the γ-ray which is subject to errors of the same order of magnitude, viz. 0·1°. It follows that the backward projection of the mean line of the pair will not, in general, pass through the centre of the parent disintegration, but at a distance r from it (Figure 37).

A simple analysis allows the expected distribution in the values of r, for any assumed values of the energy of the neutral mesons, and of their mean life-time, to be calculated. It is a satisfactory feature of the method that for $B>2$ the distributions are almost independent of the values of B. It follows that the calculated distributions, for the case in which the kinetic energy of the neutral mesons is no longer assumed to be constant, are insensitive to the assumed energy spectrum of the neutral particles, provided most of them have an energy greater than 100 Mev. This is due to the fact that although the emitted γ-radiation tends to be more closely collimated about the line of motion of the parent mesons as B increases—a factor which, if acting alone, would tend to reduce the values

of r—the effect is compensated by the relativistic extension of the time scale of the moving particles with increase in the value of B.

The actual values of r observed are shown in Figure 38 and prove that if the γ-radiation is attributed to the decay of neutral mesons the life-time of the parent particles must be less than 5×10^{-14} sec. The results suggest that the observed distributions are due to a life-time of about $2{\cdot}5 \times 10^{-14}$ sec., and are not seriously affected by errors in measurement. If further results confirm that the

Figure 37. Principle of a method of determining the lifetime of neutral mesons.

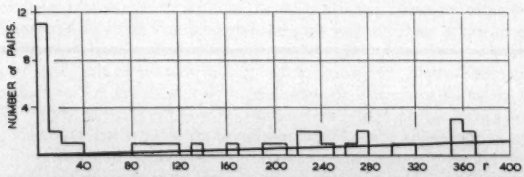


Figure 38. Observations to determine the life-time of neutral mesons. The quantity r is the orthogonal distance from a 'star' on to the bisector of the lines of motion of the electrons of a 'pair'. The concentration of events for which r is less than 40 μ, is due to 'related pairs'. Such events prove the origin of γ-radiation in nuclear explosions and allow its intensity, per star, to be determined.

spread in the values of r are indeed due to a finite life-time of this order of magnitude, the experiments may be taken to establish the independent existence of the neutral mesons. From the frequency of occurrence of pairs of electrons in the neighbourhood of stars, Carlson *et al.* have shown that the ratio of the number of neutral to charged mesons emitted in nuclear explosions is

$$\mathcal{N}(\pi^0)/\mathcal{N}(\pi^+) = 0{\cdot}45 \pm 0{\cdot}1,$$

a result in good agreement with theoretical anticipations.

8 (vi). *Spin and Parity of the π^0-Particles*

A brief reference has been made above, § 8 (ii), to the experiments of Panofsky *et al.* (1950) on the capture of π^--particles in hydrogen gas, which lead to the emission of γ-radiation. The mode of decay into two photons of the neutral meson which is the immediate result of the process of capture is inconsistent with the view that the π^0-particles are of spin 1. It is therefore reasonable to suppose that, like the charged particles, they are of spin 0. Further, the characteristics of the process of capture, in which the resulting neutron and π^0-particle are emitted with small kinetic energy, appear to indicate that the parities of the negative and neutral π-mesons are equal.

It is an interesting and peculiar feature of the results of these workers that no photons have been observed in similar experiments in which the hydrogen gas was replaced by hydrogen compounds. When π^--particles are arrested in such compounds—in lithium hydride for example—it is to be expected that many of them will be captured by lithium, and thence will eventually react with the nuclei and produce nuclear explosions. Some, however, should be captured by hydrogen, and these would be expected, at first sight, to lead to the emission of γ-radiation. The authors suggest that the absence of this radiation is due to a migration, from its original point in the crystal lattice, of the neutral complex formed by a proton and a π^--particle: a migration which takes place before the π^--particle has fallen to the state of lowest energy round the proton. As a result of this diffusion, the complex collides with neighbouring heavy atoms. In such collisions there is a large probability of the π^--particle being captured from the proton by one of the heavy atoms and then by its nucleus. The assumption of the existence of such a neutral complex between a proton and a negative meson was made by Frank (1947) in an attempt to find an alternative explanation of the π–μ events other than that of a spontaneous decay of the π-particle.

§ 9. EVIDENCE FOR THE EXISTENCE OF MESONS OF OTHER TYPES

During the past five years, physicists in a number of different laboratories have presented evidence which has been interpreted in terms of the existence of particles intermediate in mass between π-mesons and protons. Although we may be dealing with several different types, it will be convenient to refer to these particles as τ-mesons. In several instances the evidence has been based on a few isolated events observed in Wilson chambers or photographic plates; and, until recently, it has been an unsatisfactory feature of the observations that no two of them appeared to correspond to the same type of process. Opinion on the question of the existence of τ-mesons has therefore vacillated. In several instances, however, recent observations have confirmed the earlier work and strong evidence for the existence of particles of mass of the order of $1{,}000\,m_e$ is now available.

9 (i). *Evidence for the Existence of τ-Mesons*

(*a*) The first observations were those of Leprince-Ringuet and Lheritier (1944) who observed the collision of a fast cosmic-ray particle with an electron in a Wilson chamber operated in a magnetic field. From an analysis of the event they were led to suggest that the mass of the incident particle, which carried a positive charge, was about $1{,}000\,m_e$.

(*b*) Rochester and Butler (1947) observed two events in a Wilson chamber, in the first of which two particles appeared to originate in the gas of the chamber, the directions of motion being inclined at about 65°. The two particles, if they were moving away from the apparent point of origin, were of opposite sign. From the curvature of the tracks in the magnetic field, and the specific ionization of the particles as deduced from the density of droplets along the trajectory, it was concluded that the two secondary particles were of mass approximately $300\,m_e$, and it was suggested that they originated by the spontaneous decay 'in flight' of a neutral particles of mass not less than about $800\,m_e$.

In the second of the events a sudden deviation in the direction of motion of a particle in the gas of the chamber was observed. It was very difficult to account for this deviation in terms of nuclear scattering, and Rochester and Butler showed that the observed characteristics were consistent with the assumption that the event represented the decay in flight of a charged particle, of mass about $1{,}000\,m_e$, with the emission of a charged secondary particle and one or more neutral particles.

(*c*) Bradt and Peters, from a study of the grain-density in the tracks of a large number of particles in photographic plates exposed at great altitudes, suggested tentatively, at the Bristol Conference in 1948, that particles of mass approximately $700\,m_e$ are sometimes ejected with small kinetic energy from nuclear explosions. At the time the results were given with great reserve because of the possibility of pre-exposure of the plates and fading of the latent image. Further, it was an unsatisfactory feature of the observations that the particles showed no evidence for spontaneous decay or interaction with nuclei at the end of their range. The authors subsequently withdrew their suggestion as a result of further experiments, attributing their observations to a pre-exposure of the plates and fading of the latent image.

Other experimenters have made similar observations with varying results. Von Friesen and his colleagues (private communication) in Lund find no evidence for particles of mass in the interval from 500–$1{,}500\,m_e$ in photographic plates exposed at high altitudes by means of free balloons. Brown and Fowler obtained a similar result in an examination of plates exposed at an altitude of 11,000 ft. on the Jungfraujoch. On the other hand, Alichanian, Alichanow and their colleagues, using the same method, claim to have obtained results which support their conclusions regarding the existence of 'varitrons'.

(*d*) Leprince-Ringuet and his colleagues (1948) have observed an event in a photographic plate with the following features: A particle produces a track with characteristics which suggest that its mass is less than that of a proton. At the end of its range it appears to lead to a disintegration from which a π^--particle emerges, and which produces a second disintegration. From the total release of energy in the first disintegration, and assuming that the initial track is indeed that of the parent particle which produced the event, it follows that the rest-mass of this particle is approximately $700\,m_e$.

(*e*) Brown *et al.* (1949 a) have observed an event in an 'electron-sensitive' emulsion which appears to correspond to the spontaneous decay, at the end of its range, of a particle with a mass of approximately $1{,}000\,m_e$, into three π-particles. The interpretation is based on the fact that the total momentum of the three product particles is equal to zero, within the limits of error of the observations, and the total release of energy corresponds to that released by the transformation of the rest-mass of the parent particle. One of the ejected particles is brought to rest in

the emulsion and produces a nuclear disintegration so that it can be identified with confidence as a π^--particle. A second particle, of much greater energy, can be identified as a π-meson by measurements of scattering and grain-density along the track.

After a considerable lapse of time, during which opinion in the subject has fluctuated, new experimental evidence has been obtained which gives strong support for the observations of Rochester and Butler. Thus Seriff *et al.* (1950), in experiments with an expansion chamber operated at about 11,000 ft. have observed more than twenty examples of forked tracks, which appear to originate by the decay of a neutral particle τ^0 similar to the event observed by Rochester and Butler. In some cases the decay of the τ^0-meson is accompanied by a nuclear disintegration in a lead plate in the chamber. It is then reasonable to assume that the two events are associated, and that the neutral τ^0-meson was created in the disintegration. The authors state that, if this is so, the characteristics of the forks indicate that one or more neutral particles are rarely emitted in the spontaneous decay of the parent neutral particle.

Anderson and his colleagues have also reported a few examples of the second type of τ-meson described by Rochester and Butler which has been interpreted to correspond to the decay of a charged meson of mass approximately $700\,m_e$ into a charged meson of smaller mass and one or more neutral particles.

The observations of Brown *et al.* have also received support from recent work by Harding (1950) on photographic plates exposed under 3 m. of ice at the Jungfraujoch. Harding has found two events in each of which a particle of mass approximately $1{,}000\,m_e$ appears to stop in the emulsion and to decay spontaneously into three charged particles which are emitted in directions which are co-planar. A detailed analysis shows that the features of the tracks are consistent with spontaneous decay into three charged π-particles. It is an important feature of Harding's result that in the one case in which it was possible to measure the total kinetic energy of the secondary particles, the result obtained was closely similar to that of Brown *et al.*, viz. $\sim$80 MeV. If the secondary particles are indeed all π-particles, this corresponds to a mass of the parent τ-meson of the order of $985 \pm 20\,m_e$. Although it is not certain that the τ-mesons are all of the same kind, the limited evidence available is consistent with the assumption that some of the processes of decay of the three types can be represented by the equations

(1) $\tau^0 \rightarrow \pi^+ + \pi^- + \pi^0$ Rochester and Butler, type 1.
or $\tau^0 \rightarrow \pi^+ + \pi^-$

(2) $\tau^+ \rightarrow \pi^+ + \pi^0 + \pi^0$ Rochester and Butler, type 2.
or $\tau^+ \rightarrow \pi^+ + \pi^0$

(3) $\tau^+ \rightarrow \pi^+ + \pi^+ + \pi^-$ Brown *et al.*

For the present, however, the first four of these equations must be regarded as highly speculative.

The work of several experimenters on the nature of the mesons created in the nuclear interactions of protons and α-particles of great energy in the cosmic radiation show that any τ-mesons sufficiently stable to be brought to rest in solid materials are produced rarely compared with the π-mesons. The three events of the type observed by Brown *et al.*, and by Harding, have all been observed at an altitude of 11,000 ft., under considerable thicknesses of materials, and it appears

possible that they are commonly produced only with great kinetic energy. In this case it would be possible to observe them only in the presence around the photographic plates of considerable thicknesses of absorbing materials. In the events described by Rochester and Butler and by Seriff *et al.*, the life-time of the particles can hardly be less than 10^{-10} sec., and of those described by Brown *et al.*, not less than 10^{-12} sec.

9 (ii). *Varitrons*

The experiments of Alichanian *et al.* (1948, 49), to which reference has already been made, have been carried out by means of an apparatus operated at an altitude of 10,000 ft., and of the type shown schematically in Figure 39. Charged particles of the cosmic radiation of which the directions of motion are inclined at only small

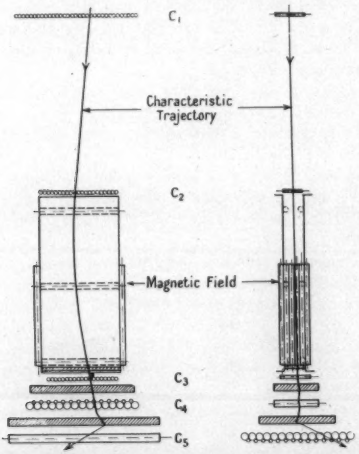


Figure 39. Apparatus of Alichanian, Alichanov and Weissenberg (1948), for the determination of the mass of charged particles of the cosmic radiation.

angles to the vertical, pass through the arrays of counters at C_1 and C_2. By observing which of the individual counters are discharged, the direction of motion of the particle is determined. After emerging from tray C_2 a particle passes between the poles of a permanent magnet, giving a field strength of 7,000 gauss, and is deflected so that it leads to the discharge of a counter in tray C_3 out of line with its original direction of motion. The deviation in the trajectory can thus be observed and the momentum of the particle determined within limits due to the finite size of the counters.

After passing through the magnetic field the particle passes through sheets of lead between which there are further arrays of counters. It is thus possible to determine the range of the particles from observations on the thickness of lead which they are able to traverse, the uncertainty in the determination being limited by the thickness of the lead plates. In principle the combination of the two observations allows the rest-mass of the particle to be determined, provided the

particle is stopped in its passage through the lead solely as a result of loss of energy by ionization (see § 3 (iv)).

As a result of the finite thickness of the lead plates, the values of the deflections of particles of the same mass which are arrested in a given layer of lead, will be distributed in the way indicated in Figure 40 (*a*). The authors interpret their experimental observations as giving evidence for the existence of such edges in their distribution curves, and conclude that many types of charged mesons exist

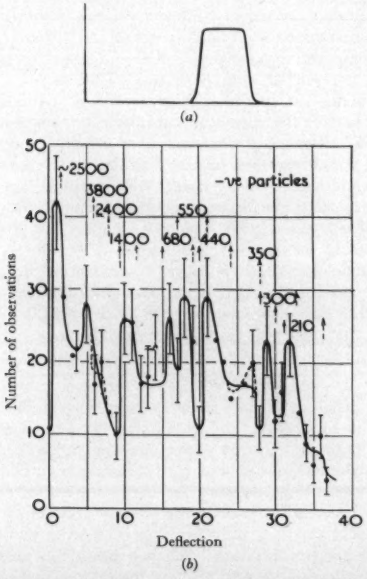


Figure 40. Characteristic result of Alichanian *et al.* (1948).

which they refer to as 'varitrons', a name chosen to indicate the diversity of the values of the rest-mass of the particles. A typical example of one of their observed distributions is shown in Figure 40 (*b*).

A preliminary report has recently been given by Brode (1949) of experiments similar in principle to those described above, in which the momentum of a particle is deduced from the curvature of its trajectory in a Wilson chamber operated in a magnetic field. Further, the arrest of the particle is studied as it passes through lead plates arranged in a Wilson chamber (see Figure 9 (*b*)). Apart from μ^+ and μ^- mesons and protons, Brode observed six particles, all positively charged, of which the mass appeared to be about $1{,}000\,m_e$. Recently, however, he has concluded that these six particles were in fact protons.

The results of the Soviet experiments appear to be in contradiction with those of Franzinetti (1950), Fowler (1950), and other workers whose experiments have

certain advantages already discussed. In experiments with counters it is difficult to be certain that processes, other than those which alone are assumed to be operative, are not in fact making contributions to the discharges of the counters. It is possible, for example, that in Alichanian's experiments, some of the observations are due to particles moving vertically upwards. Such particles are sometimes generated in the nuclear collisions of fast protons and neutrons entering the surface layers of the earth, and are known to exist in appreciable intensity (Camerini *et al.* 1948).

As a second possible source of confusion in the Soviet experiments, the particles of the dominant downward-moving stream will not invariably be brought to rest by loss of energy through ionization alone. Some of them will be stopped catastrophically as a result of making nuclear collisions. In other cases the electrons emitted during the spontaneous decay of a μ^+-particle may discharge counters at a lower level in the apparatus and thus lead to an over-estimate of the range of the particle. Because of these and other possible processes, the limited statistical weight of the observations obtained hitherto, and the failure of workers in other laboratories to obtain similar results with improved apparatus, it seems reasonable to suggest that the evidence should for the time being be treated with great reserve (see also Daudin 1950).

The experiments of Brode allow a much closer study to be made of the trajectories of the particles and of the physical processes occurring in the apparatus. It is thus possible to exclude most of the ambiguities present in the counter experiments. In this case also, however, the particles are commonly stopped in the lead plates so that some secondary processes which they may produce escape observation.

9 (iii). *Summary of Evidence*

The question of the existence of other types of mesons, in addition to the π- and μ-particles is of decisive importance for the development of nuclear physics. The evidence summarized above strongly suggests that other types sufficiently stable to be brought to rest in solid materials without having a high probability of decaying in flight, do exist, but that they are rare compared with the π-particles. It appears to be a matter of the greatest importance to improve the precision and statistical weight of the observations, in order that the existence of the particles may be put beyond doubt and their properties established, for the present difficulties in developing a satisfactory meson theory may be due to lack of decisive information. It was pointed out at the beginning of the article that the discovery of the π-meson was made very soon after a sufficient increase in the sensitivity of the photographic method. The possibility must therefore be borne in mind of the existence of other types of particles even more elusive, and with even shorter life-times than the π- and τ-mesons.

§10. PRODUCTION OF MESONS IN THE ATMOSPHERE

As a result of the discoveries of the past four years, it now appears to be possible to give a coherent account of the main processes associated with the passage of cosmic radiation which take place in the atmosphere. It is hardly likely that our present picture will stand in its entirety: it has still to be supported by numerical calculation, many points of detail remain to be elucidated and, in a field of study which has proved rich in yielding fundamental new phenomena,

we may still expect surprises. Nevertheless, it seems reasonable to suppose that many of the most important features of our present views are essentially correct.

10 (i). *The Primary Cosmic Radiation*

The discoveries of Freier *et al.* (1948) and of Bradt and Peters (1948) show that in addition to the energetic protons which form the most numerous of the incoming particles, the primary cosmic radiation incident in the high atmosphere contains a large fraction of heavier nuclei. The track of one of these particles, observed in photographic plates exposed at an altitude greater than 100,000 ft., is shown in Plate X. Preliminary results by Bradt and Peters (1950) on the distribution in mass of the particles of the primary radiation, shown in Figure 41, indicate that about 55% of the incoming matter arrives in the form of nuclei more massive than protons.

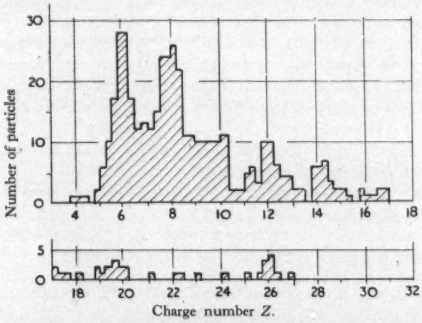


Figure 41. Distribution in charge of the heavy particles of the primary cosmic radiation. (Bradt and Peters 1950.)

It has recently been pointed out by Bradt and Peters that the mass spectrum of the primary radiation shows notable gaps, the relative abundance of Li, Be, B and F being particularly low in comparison with neighbouring elements. Apart from its intrinsic interest, this result proves that most of the primary particles have not made a nuclear collision in their journey through space. For otherwise any detailed structure in the form of the mass-spectrum would have been obliterated—see for example Plate XII. Any theory of the origin of the cosmic radiation must explain these features of the primary component.

Because of the magnetic field of the earth, only particles with a certain minimum energy can approach the atmosphere at points with a given magnetic latitude. At latitude 50° N. for example, the protons must have a minimum energy of 1,500 MeV. The energy of heavier particles is limited by similar considerations and depends essentially on the value of Ze/M, where Ze is the charge and M the mass. It is a sufficient approximation, for elements of medium atomic weight, to assume that Ze/M is equal to one half the corresponding value for the proton, and with this assumption the minimum energy of the heavy particles is equal to 500 MeV. *per nucleon.* The minimum energy of the incoming nuclei therefore increases with their mass. This has the effect that if we confine attention to

particles with energies above a certain value, say 10^{11} ev., the heavier particles make a much larger contribution than in the case of the total incoming stream. A more detailed study will depend on the determination of the distribution in energy of the incoming particles, a second point of great interest in its bearing on the problem of the origin of the cosmic radiation.

10 (ii). *Disintegrations of Heavy Primary Particles*

Because of the large charge carried by the heavy particles, their rate of loss of energy by ionization is very great, and they have a large cross section for interaction with nuclei. As a result they rarely penetrate to depths in the atmosphere below 60,000 ft. Further, the high probability of making a nuclear collision has the consequence that, unlike the particle of which the track is shown in Plate X, relatively few of them reach the end of their range before being disrupted.

The collisions of energetic heavy particles with other nuclei result in a large variety of types of nuclear explosions, (Bradt and Peters 1949) of which a few examples are reproduced in Plates IX and XII. In some cases the incident nucleus appears to be almost completely disrupted into its component nucleons. In others, presumably due to a more distant collision, the disintegration is less complete—see Plate XII. The collisions sometimes lead to the production of showers of π-mesons similar to those produced by neutrons, protons and α-particles moving at relativistic velocities, and because of the large numbers of nucleons involved, the number of mesons in such a shower may be very great.

As a result of processes of this type, showers of fast nucleons moving through the atmosphere contemporaneously, and with energies in the relativistic region, are produced in addition to the primary protons. This stage of the development of the secondary processes is represented in Figure 42 (*a*).

10 (iii). *Origin of the 'Hard' and 'Soft' Components*

The fast individual nucleons can penetrate to much greater depths in the atmosphere than the heavy nuclei. Due, however, to nuclear collisions, their intensity decreases exponentially with the mass of atmosphere penetrated. The particles interact with nuclei and produce disintegrations and showers of fast charged π-mesons most of which are created with an energy less than 1,000 Mev. Because of their short life-time, these π-particles, when moving in the atmosphere, commonly decay in flight after moving distances less than 100 metres from their points of creation. By transforming into μ-mesons they give rise to the greater part of the 'hard' component of the cosmic radiation.

The nuclear explosions are also accompanied by the emission of neutral π-mesons of very short life-time. The spontaneous decay of these particles into γ-radiation, and the subsequent production of the cascade showers of electrons and photons by 'pair-production' and 'bremsstrahlung', allows us to account for most of the 'soft' radiation. Both the 'hard' and 'soft' components therefore appear to arise in nuclear processes of the same type. The distributions in energy of the charged and neutral π-particles are closely similar and the number of neutral particles is approximately half that of the charged mesons. The values of total energy carried by the hard and soft components are therefore closely similar, a result in accord with the observations made by Rossi 1948. These processes are represented schematically in Figure 42 (*b*).

We may assume that many of the particles in a group of fast nucleons, arising from the disintegration of an original heavy 'primary' nucleus of great energy,

will be able to create ' showers ' of π-mesons. If so, we can account for the production of the ' extensive penetrating showers '—groups of many hundreds of μ-mesons accompanied by soft radiation. An original ' primary ' nucleus may, in a collision, give rise to many nucleons. The showers of penetrating particles, subsequently produced in the atmosphere by many of these nucleons, will appear to be contemporaneous, for the variation in their instants of arrival at a given depth will be undetectable with the apparatus commonly employed. The extensive penetrating showers may therefore be attributed both to primary heavy nuclei and to primary protons of sufficiently great energy. It will be possible to estimate the relative importance of these two processes when accurate measurements of the distribution in charge and energy of the primary radiation have been made.

There is some evidence that the direct production of charged and neutral π-particles in the nuclear explosions of great energy is accompanied by the creation

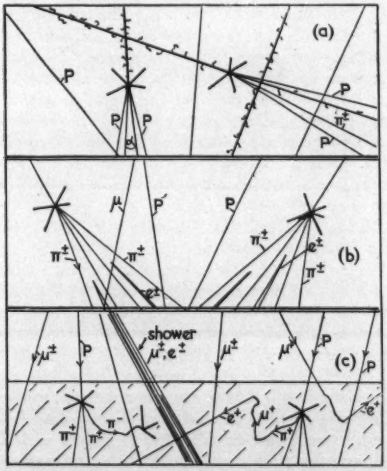


Figure 42. Schematic representation of processes occurring in the atmosphere.

of energetic τ-mesons with a frequency of the order of 1% of that of the π-particles. These τ-mesons, at least in some cases, appear to decay spontaneously into π-particles with a lifetime of approximately 10^{-10} sec.

Because of their very weak interaction with nucleons, those of the μ-mesons with very great energy can penetrate to great depths underground. They can then occasionally interact with nuclei to produce nuclear explosions accompanied by the emission of showers, possibly through the electromagnetic interaction with protons in the nuclei.

The downward flux of charged particles and photons at any depth in the atmosphere is accompanied by a flux of ' neutrinos '—neutral particles, of small rest-mass of which the fate is, at present, unknown.

§11. OUTSTANDING PROBLEMS AND PERSPECTIVES

The experiments carried out during the past three years have established the fact that the π-mesons have properties closely similar to those visualized for the ' heavy quanta '. The particles are created in nucleon–nucleon collisions; they interact with the nucleus of every, or almost every, nucleon with which they collide ; when brought to rest in solid materials they commonly interact with the nuclei, even of hydrogen, before they have had time to decay, and this in spite of the short life-time of the particles, viz. $\sim 10^{-8}$ sec. At one time it seemed reasonable to hope, therefore, that it would be possible to modify the original conceptions of Yukawa, and to establish a satisfactory meson theory which would enable a consistent account to be given of the particles and of their relationship to the forces of nuclear cohesion.

With this end in view a great many variations of Yukawa's theory have been developed, in the mathematical formulations of which the meson field has been treated, alternatively, as a scalar quantity (as in the original theory), as a vector, as a pseudo-scalar (spin-dependent terms being added to the terms in the original ' scalar theory ' for the forces between two nucleons), and in many other ways. Recent experiments on the generation of π-mesons by γ-rays, on the capture of π^--particles in gaseous hydrogen, and on the mode of decay of π^0-particles into two γ-rays, provide important information which proves that many of the varieties of meson theory are inconsistent with the observations.

In spite of the progress represented by the new discoveries, the question remains, however, whether, in spite of the great stimulus which they have provided for the progress of theory and experiment, the basic features of Yukawa's original ideas are essentially correct. Even if it is a valuable analogy to identify the π-mesons with the ' heavy quanta ' of the nuclear field, there appears to be no place in any formalism, developed hitherto, for the μ-mesons. These particles have a mass of $215\, m_e$; they have a very weak interaction with nucleons, so that they are able to penetrate hundreds of nuclei without disintegrating them ; they have half integral spin and decay with the emission of an electron and, probably, two neutrinos; they have a life-time of $2{\cdot}1 \times 10^{-6}$ sec., and the negative particles when arrested in elements with Z greater than 15 interact with the nuclei before decay and lead to a nuclear transmutation with the emission of a neutrino and a neutron. The great inadequacy of our present theoretical ideas is well illustrated by the fact that there has scarcely been an attempt to account for the existence of these particles and the details of their properties.

These difficulties were already present at a time, a few months ago, when it was commonly believed that we had not to face the additional complication represented by the existence of other types of particles. The recent additional evidence for the existence of τ-mesons strongly suggests, however, that the situation is much more complex than had commonly been visualized : that particles of mass about $1{,}000\, m_e$ exist, both charged and uncharged, which decay, at least in some cases, with the emission of π-particles. The most important features of the most common of the different types of particles, of which any theory must attempt to explain the existence and properties, are summarized in Table 4.

We have no reason to believe that the particles included in Table 4 exhaust the list of those most transient forms of matter which we refer to as mesons. We know for example that the neutral π-particles have a life-time less than 5×10^{-14} sec. It would be very difficult to establish the independent existence

of more transient forms of matter, either charged or uncharged, with the present technical resources at our disposal. Without considering such possibilities, it is clear that any theoretical treatment must provide an explanation of the variety of mesons and their properties displayed in Table 4, a variety which appears to us, in the present stage of the development of the subject, as a succession of arbitrary and largely unrelated facts, but which must represent an underlying unity which at present escapes us.

"Experimental science", says Clerk Maxwell, "is continually revealing to us new features of natural processes, and we are thus compelled to search for new forms of thought appropriate to those features". It is possible that decisive

Table 4. Tentative List of different Types of Mesons and their Properties

Type	Reference	Mass in m_e	Mean life-time *in vacuo*	Spin	Mode of decay
μ^+	Anderson (1938)	212 ± 4	$2{\cdot}15\times10^{-6}$ sec.	$(\frac{1}{2})$	$\mu^+\rightarrow e^+ + \nu + \nu$
μ^-		212 ± 4	$2{\cdot}15\times10^{-6}$ sec.	$(\frac{1}{2})$	$\mu^-\rightarrow e^- + \nu + \nu$ $\mu^- + X_A^Z \rightarrow X_{A-1}^{Z-1} + N + \nu$ (in heavy materials)
π^+	Lattes *et al.* (1947)	276 ± 4	$1{\cdot}6\times10^{-8}$ sec.	(0)	$\pi^+\rightarrow\mu^+ + \nu$
π^-		278 ± 4	$1{\cdot}0\times10^{-8}$ sec.	(0)	$\pi^-\rightarrow\mu^- + \nu$ $\pi^- + X_A^Z \rightarrow$ star $\pi^- + P \rightarrow \pi^0 + N$ $\pi^- + P \rightarrow h\nu + N$ (in hydrogen)
π^0	Bjorkland *et al.* (1950)	~ 270	$<5\times10^{-14}$ sec.	(0)	$\pi^0\rightarrow 2h\nu$
τ^+	Rochester and Butler (1948)	~ 900		?	$(\tau^\pm\rightarrow\pi^\pm + \pi^0)$? $(\tau^\pm\rightarrow\pi^\pm + \pi^0 + \pi^0)$?
τ^0	Rochester and Butler (1948)	~ 900	$\sim(2\times10^{-10}$ sec.)	?	$(\tau^0\rightarrow\pi^+ + \pi^-)$? $(\tau^0\rightarrow\pi^+ + \pi^- + \pi^0)$?
τ^+	Brown *et al.* (1950)	~ 1000	$>10^{-13}$ sec.	(0)	$(\tau^+\rightarrow\pi^+ + \pi^+ + \pi^-)$

ν represents any particle of small rest-mass not a proton.

Features for which the evidence is not yet conclusive, or which are speculative, are shown in brackets. The difference in the values of the life-times of the π^+ and π^--particles is not established.

information which could provide the stimulus for a radically new approach to the problem of the mesons remains to be discovered. Great interest therefore attaches to the new proton-synchrotrons under construction. The particles generated by the synchro-cyclotrons now in operation are not sufficiently energetic to be able, from the point of view of the conservation laws, to create particles with a mass of the order of 1,000 m_e. The Birmingham proton-synchrotron is designed, however, to produce protons of energy greater than 1,000 MEV., and these may be able to produce τ-mesons artificially. The machines, based on similar principles under construction at Brockhaven and at Berkeley, are designed to accelerate protons to energies of 3,000 MEV. and 6,000 MEV., respectively.

Even when the new machines have been brought successfully into operation, however, it will still be necessary to turn to natural sources in order to study the nuclear transmutations produced by particles of the greatest energy. Experiments with Wilson chambers operated in magnetic fields, and the development of new methods of determining the energy of fast particles in photographic emulsions, have recently made it possible to make a detailed study of the disintegrations produced by protons and α-particles with energies in the interval from 10^9 to 10^{11} ev., and occasionally by particles of even greater energy.

As a result of these developments there is to-day no line of division between nuclear physics and the study of cosmic radiation. The latter can be regarded as nuclear physics of the extreme high energy region.

In view of these technical developments, and of the very great resources in men and materials which have been concentrated in the study of these high-energy processes, it appears to be certain that, given peace, the coming years will see further rapid advances in our knowledge of the mesons, those particles which, when free, are ephemeral, but which appear to play a decisive rôle in ensuring the stability of the nuclei and, therefore, of the order of the material world as we know it.

ACKNOWLEDGMENTS

In preparing this report I have been greatly indebted to many of my colleagues for discussions on points of detail, and especially to Mr. H. Muirhead, Mr. U. Camerini and Dr. C. Franzinetti. I have made frequent reference to Professor Rossi's reports on cosmic radiation, to Professor Jánossy's *Cosmic Radiation* and to Dr. H. Bradner's *Review of Work on Artificially Produced Mesons*. The photo-micrographs are largely the work of Mr. R. Gattiker and Miss P. Dyer, to whom my particular thanks are due.

REFERENCES

The following list does not include references to many important contributions to the subject for which space was not available in the text.

ADELMAN, F. L., and JONES, S. B., 1949, *Phys. Rev.*, **75,** 1468 (A).

ALICHANYAN, A. I., ALICHANOV, A. I., MOROZOV, V., and KHRIMIAN, A., 1948, *Acad. Sci. Rep.*, **61,** 35, and earlier references.

ALICHANYAN, A. I., ALICHANOV, A. I., and WEISSENBERG, A., 1948, *J. Exp. Theor. Phys. U.S.S.R.*, **18,** 301.

ALVAREZ, L. W., LONGACRE, A., OGREN, V. G., and THOMAS, R. E., 1950, *Phys. Rev.*, **77,** 752 (A).

ANDERSON, C. D., ADAMS, R. V., LLOYD, P. E., and RAU, R. R., 1947, *Phys. Rev.*, **72,** 724.

ANDERSON, C. D., and NEDDERMEYER, S., 1937, *Phys. Rev.*, **51,** 884; 1938, *Phys. Rev.*, **54,** 88.

AUGER, P., EHRENFEST, P., FREON, A., and FOURNIER, M., 1937, *C. R. Acad. Sci., Paris*, **204,** 257.

AUGER, P., MAZE, R., and CHAMINADE, R., 1941, *C. R. Acad. Sci., Paris*, **213,** 381.

BARBOUR, I., 1949, *Phys. Rev.*, **76,** 320; 1950, *Ibid.*, **77,** 751, **78,** 518.

BARKAS, W. H., 1949, *Phys. Rev.*, **75,** 1467 (A).

BARKAS, W. H., BISHOP, A. S., BRADNER, H., and SMITH, F. M., 1950, *Phys. Rev.*, **77,** 462.

BARKAS, W. H., BISHOP, A. S., GARDNER, E., and LATTES, C. M. G., 1950, *Phys. Rev.* (in press); see also 1949, *Phys. Rev.*, **75,** 1467, 8, and *Berkeley Report*, October 1949.

BERRIMAN R., 1948, *Nature, Lond.*, **162,** 992.

BISHOP, A. S., BRADNER, H., and SMITH, F. M., 1949, *Phys. Rev.*, **76,** 588 (A).

BJORKLAND, R., CRANDALL, W. E., MOYER, B. J., and YORK, H. F., 1950, *Phys. Rev.*, **77,** 213.

BLACKETT, P. M. S., and OCCHIALINI, G. P. S., 1933, *Proc. Roy. Soc.* A, **139,** 669; 1936, *Ibid.*, **154,** 573.

BLACKETT, P. M. S., and WILSON, J. G., 1938, *Proc. Roy. Soc.* A, **165,** 209.

BOWKER, J. K., 1950, *Phys. Rev.*, **78,** 87 (A).

BRADNER, H., 1949, *Review of Work on Artificially Produced Mesons*, Berkeley, October, 1949.

BRADNER, H., and JONES, S. B., 1950, *Phys. Rev.*, **78,** 90.

BRADT, H. L., KAPLAN, J., and PETERS, B., 1950, *Helv. Phys. Acta*, **23,** 24.

BRADT, H. L., and PETERS, B., 1948, *Phys. Rev.*, **74,** 1828; 1949, *Ibid.*, **75,** 1779; 1950, *Ibid.*, **77,** 54.

BRODE, I. R. B., 1949, *Rev. Mod. Phys.*, **21,** 37, *Report Como Conf.*

BROWN, R. H., CAMERINI, U., FOWLER, P. H., HEITLER, H., KING, D. T., and POWELL, C. F., 1949 b, *Phil. Mag.*, **40,** 862.

BROWN, R. H., CAMERINI, U., FOWLER, P. H., MUIRHEAD, H., POWELL, C. F., and RITSON, D., 1949 a, *Nature, Lond.*, **163,** 47, 82.

BROWN, W. W., and MCKAY, A. S., 1950, *Phys. Rev.*, **77**, 342.
BRUECKNER, K. A., 1949, *Phys. Rev.*, **78**, 84 (A).
BUTLER, C. C., ROSSER, W. G. V., and BARKER, K. H., 1950, *Proc. Phys. Soc.* A, **63**, 145.
CAMERINI, U., COOR, T., DAVIES, J. H., FOWLER, P. H., LOCK, W. O., MUIRHEAD, H., and TOBIN, N., 1949, *Phil. Mag.*, **40**, 1073.
CAMERINI, U., FOWLER, P. H., LOCK, W. O., and MUIRHEAD, H., 1950, *Phil. Mag.*, **41**, 413.
CAMERINI, U., MUIRHEAD, H., POWELL, C. F., and RITSON, D. M., 1948, *Nature, Lond.*, **162**, 433.
CARLSON, A. G., HOOPER, J. E., and KING, D. T., 1950, *Phil. Mag.*, **41**, 701.
CARTWRIGHT, W. F., RICHMAN, C., WHITEHEAD, M. M., and WILCOX, H. A., 1950, *Phys. Rev.*, **78**, 823.
CHAMINADE, R., FREON, A., and MAZE, R., 1944, *C. R. Acad. Sci., Paris*, **218**, 402.
CHANG, W. Y., 1949, *Rev. Mod. Phys.*, **21**, 133; also pp. 153, 166.
CONVERSI, M., PANCINI, E., and PICCIONO, O., 1947, *Phys. Rev.*, **71**, 209.
COSYNS, M., DILWORTH, C. C., OCCHIALINI, G. P. S., SHOENBERG, M., and PAGE, N., 1949, *Proc. Phys. Soc.* A, **62**, 801.
DAVIES, J. H., LOCK, W. O., and MUIRHEAD, H., 1949, *Phil. Mag.*, **40**, 1250.
DAUDIN, J., 1950, *J. Phys. Radium*, **11**, 25 (D).
FERMI, E., TELLER, E., and WEISSKOPF, V., 1947, *Phys. Rev.*, **71**, 314.
FERRETTI, B., and GALLONE, S., 1950, *Phys. Rev.*, **77**, 153.
FLINT, H. T., 1939, *Rep. Prog. Phys.*, **5**, 402 (London : Physical Society).
FOWLER, E. C., COOL, R. L., and STREET, J. C., 1948, *Phys. Rev.*, **74**, 101.
FOWLER, P. H., 1950, *Phil. Mag.*, **41**, 169.
FRANK, F. C., 1947, *Nature, Lond.*, **160**, 525.
FRANZINETTI, C., 1949, *Report Como Conf.*; 1950, *Phil. Mag.*, **41**, 86.
FREIER, P., LOFGREN, E. J., NEY, E. P., OPPENHEIMER, F., BRADT, H. L., and PETERS, B., 1947, *Phys. Rev.*, **74**, 213.
FRETTER, W., 1946, *Phys. Rev.*, **70**, 625; 1949, *Ibid.*, **76**, 511.
FRÖHLICH, H., 1947, *Nature, Lond.*, **160**, 255.
FRÖHLICH, H., HEITLER, W., and KEMMER, N., 1938, *Proc. Roy. Soc.* A, **166**, 154.
GARDNER, E., and LATTES, C. M. G., 1948, *Science*, **107**, 270.
GEORGE, E. P., and EVANS, J., 1950, *Proc. Phys. Soc.* A (in press).
GEORGE, E. P., and JASON, A. C., 1949, *Cosmic Radiation* (Colston Papers : London : Butterworth), p. 31.
GOLDSCHMIDT-CLERMONT, Y., 1950, *Nuovo Cim.*, **7**, 1.
GOLDSCHMIDT-CLERMONT, Y., KING, D. T., MURIHEAD, H., and RITSON, D. M., 1948, *Proc. Phys. Soc.*, **61**, 138.
HALPERN, O., and HALL, H., 1950, *Phys. Rev.*, **78**, 320 (A).
HARDING, J. B., 1950, *Phil. Mag.*, **41**, 405.
HARDING, J. B., and PERKINS, D. H., 1949, *Nature, Lond.*, **164**, 285.
HEIDMANN, J., and LEPRINCE-RINGUET, L., 1948, *C. R. Acad. Sci., Paris*, **226**, 1716.
HEISENBERG, W,, 1949, *Z. Phys.*, **126**, 569.
HEITLER, W., 1939, *Rep. Prog. Phys.*, **5**, 361 (London : Physical Society).
HEITLER, W., and JÁNOSSY, L., 1949, *Proc. Phys. Soc.* A, **62**, 364, 669; see also HAMILTON, J., HEITLER, W., and PENG, H. W., 1943, *Phys. Rev.*, **64**, 78, and JÁNOSSY, L., 1943, *Phys. Rev.*, **64**, 345.
HINCKS, E. P., and PONTECORVO, B., 1948, *Phys. Rev.*, **73**, 257; see also 1949, *Ibid.*, **75**, 698; 1950, *Ibid.*, **77**, 102.
JÁNOSSY, L., 1941, *Proc. Roy. Soc.* A, **179**, 361.
JONES, S. B., and WHITE, R. S., 1948, *Phys. Rev.*, **75**, 1468; 1950, *Ibid.*, **78**, 12.
KING, D. T., 1950, *Phil. Mag.*, **41**, 165, 526.
KRAUSHAAR, W. L., THOMAS, J. E., and HENRI, V. P., 1950, *Phys. Rev.*, **78**, 486.
LANDAU, L. D., 1948, *Dokl. Akad. Nauk, SSSR*, **60**, 207.
LATTES, C. M. G., FOWLER, P. F., and CUER, P., 1947, *Proc. Phys. Soc.*, **59**, 883.
LATTES, C. M. G., MUIRHEAD, H., OCCHIALINI, G. P. S., and POWELL, C. F., 1947 a, *Nature, Lond.*, **159**, 694.
LATTES, C. M. G., OCCHIALINI, G. P. S., and POWELL, C. F., 1947 b, *Nature, Lond.*, **160**, 453, 486; 1948, *Proc. Phys. Soc.*, **61**, 173.
LEIGHTON, R. B., ANDERSON, C. D., and SERIFF, A. J., 1949, *Phys. Rev.*, **75**, 1432.
LEPRINCE-RINGUET, L., 1949, *Rep. Como Conf.*

LEPRINCE-RINGUET, L., and LHERITIER, M., 1944, *C. R. Acad. Sci., Paris*, **219,** 618.
LEPRINCE-RINGUET, L., TCHANG FONG, H., JAUNEAU, L., and MORELLET, D., 1948, *C. R. Acad. Sci., Paris*, **226,** 1897.
LEWIS, H. W., OPPENHEIMER, J. R., and WOUTHUYSEN, S. A., 1948, *Phys. Rev.*, **73,** 127.
LOVATI, A., MURA, A., SALVINI, G., and TAGLIAFERRI, G., 1950, *Phys. Rev.*, **77,** 284.
MCMILLAN, W. G., and TELLER, E., 1947, *Phys. Rev.*, **72,** 1.
MARSHAK, R. E., 1949, *Phys. Rev.*, **75,** 1136.
MARSHAK, R. E., and BETHE, H. A., 1947, *Phys. Rev.*, **72,** 506.
MARTINELLI, E. A., and PANOFSKY, W. K. H., 1950, *Phys. Rev.*, **77,** 465.
MENON, M. G. K., MUIRHEAD, H., and ROCHAT, O., 1950, *Phil. Mag.*, **41,** 583.
MICHEL, L., 1950, *Proc. Phys. Soc.* A, **63,** 514.
MÖLLER, C., 1941, *Det. Kgl. Danske Videnskabernes Selskab*, **18,** No. 6; 1946, *Report of International Conference on Fundamental Particles*, I, p. 184 (London : Physical Society).
OCCHIALINI, G. P. S., 1949, *Report Como Conf.*
OCCHIALINI, G. P. S., and POWELL, C. F., 1947, *Nature, Lond.*, **159,** 186.
O'CEALLAIGH, C., 1950, *Phil. Mag.*, **41,** 838.
PANOFSKY, W. K. H., AAMODT, L., and YORK, H. F., 1950, *Phys. Rev.*, **78,** 825.
PEIERLS, R. E., 1940, *Rep. Prog. Phys.*, **6,** 78 (London : Physical Society).
PERKINS, D. H., 1947, *Nature, Lond.*, **159,** 126; 1949, *Phil. Mag.*, **40,** 101.
PETERSON, J. M., 1948, *Phys. Rev.*, **75,** 1469 (A).
PICCIONI, O., 1948, *Phys. Rev.*, **74,** 1754; see also EPSTEIN, S. T., FINKELSTEIN, R. J., and OPPENHEIMER, J. R., 1948, *Phys. Rev.*, **73,** 1140; 1950, *Phys. Rev.*, **77,** 1, 6.
POWELL, C. F., 1949, *Report Como Conf.*
RASETTI, F., 1941, *Phys. Rev.*, **59,** 613, **60,** 198.
RETALLACK, J. G., and BRODE, R. B., 1949, *Phys. Rev.*, **75,** 1716.
RICHARDSON, J. R., 1948, *Phys. Rev.*, **74,** 1720.
RICHMAN, C., and WILCOX, H., 1950, *Phys. Rev.*, **78,** 85 (A); see also WEISSBLUTH, *Ibid.*, **78,** 86 (A).
ROCHESTER, G. D., and BUTLER, C. C., 1947, *Nature, Lond.*, **160,** 855.
ROSSI, B., 1948, *Rev. Mod. Phys.*, **20,** 537 ; 1949, *Cosmic Radiation* (Colston Papers : London : Butterworth).
ROSSI, B., and NERESON, N., 1942, *Phys. Rev.*, **62,** 417, **64,** 199.
SAKATA, S., and INOUE, T., 1946, *Prog. Theor. Phys.*, **1,** 143.
SANDS, M., 1950, *Phys. Rev.*, **77,** 180.
SARD, R. D., ITTNER, W. B., CONFORTO, A. M., and CROUCH, M. F., 1948, *Phys. Rev.*, **74,** 97.
SERBER, R., 1949, *Phys. Rev.*, **75,** 1459.
SERIFF, A. J., LEIGHTON, R. B., HSIAO, C., COWAN, E. W., and ANDERSON, C. D., 1950, *Phys. Rev.*, **78,** 290.
SMITH, F. M., GARDNER, E., and BRADNER, H., 1950, *Phys. Rev.*, **77,** 562.
STEINBERGER, J., 1948, *Phys. Rev.*, **74,** 500; 1949, *Ibid.*, **75,** 1136.
STEINBERGER, J., and BISHOP, A. S., 1950, *Phys. Rev.*, **78,** 493.
STEINBERGER, J., PANOFSKY, W. K. H., and STELLER, J., 1950, *Phys. Rev.*, **78,** 802.
TAMOR, S., 1950, *Phys. Rev.*, **77,** 412.
THOMPSON, R. W., 1948, *Phys. Rev.*, **74,** 490.
TICHO, H. K., and SCHEIN, M., 1948, *Phys. Rev.*, **73,** 81.
TIOMNO, J., WHEELER, J. A., and RAU, R. R., 1949, *Rev. Mod. Phys.*, **21,** 144.
TOMONAGA, S., and ARAKI, G., 1940, *Phys. Rev.*, **58,** 90.
VALLEY, G. E., 1947, *Phys. Rev.*, **72,** 772.
VALLEY, G. E., and ROSSI, B., 1948, *Phys. Rev.*, **73,** 177.
WANG, K., and JONES, S. B., 1948, *Phys. Rev.*, **74,** 1547.
WATAGHIN, G., SANTOS, M. D. DE S., and POMPEIA, P. A., 1940, *Phys. Rev.*, **57,** 339.
WENTZEL, G., 1949, *Phys. Rev.*, **75,** 1810.
WHEELER, J. A., 1947, *Phys. Rev.*, **71,** 320; 1949, *Rev. Mod. Phys.*, **21,** 133.
WHITE, R. S., MCMILLAN, E. M., and PETERSON, J. M., 1950, *Phys. Rev.*, **78,** 84 (A).
WILLIAMS, E. J., 1939, *Proc. Roy. Soc.* A, **169,** 531; 1940, *Phys. Rev.*, **58,** 292.
WILLIAMS, E. J., and ROBERTS, G. E., 1940, *Nature, Lond.*, **145,** 102.
YANG, C. N., 1950, *Phys. Rev.*, **77,** 242, 722.
YUKAWA, H., 1935, *Proc. Phys. Math. Soc., Japan*, **17,** 48.
ZAR, J. L., HERSHKOWITZ, J., and BEREZIN, E., 1948, *Phys. Rev.*, **74,** 111.